



ITS
Institut
Teknologi
Sepuluh Nopember

TUGAS AKHIR - KS 141501

**KLASIFIKASI TEKS PERMINTAAN INFORMASI
UNTUK APLIKASI ONLINE SHOP
MENGUNAKAN ALGORITMA SUPPORT
VECTOR MACHINE (STUDI KASUS: BENTO
SHOP)**

**DEA ANDIA RACHMAWATI
NRP 5212 100 177**

**Dosen Pembimbing I
Renny Pradina Kusumawardani, S.T., M.T**

**Dosen Pembimbing II
Radityo Prasetyanto.W, S.Kom, M.Kom**

**JURUSAN SISTEM INFORMASI
Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember
Surabaya 2016**



ITS
Institut
Teknologi
Sepuluh Nopember

FINAL PROJECT - KS 141501

***TEXT CLASSIFICATION OF INFORMATION
REQUEST FOR ONLINE SHOP USING SUPPORT
VECTOR MACHINE ALGORITHM (CASE STUDY :
BENTO SHOP)***

DEA ANDIA RACHMAWATI
NRP 5212 100 177

Supervisor I
Renny Pradina Kusumawardani, S.T., M.T

Supervisor II
Radityo Prasetyanto.W, S.Kom, M.Kom

INFORMATION SYSTEM DEPARTEMENT
Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya 2016

LEMBAR PENGESAHAN

KLASIFIKASI TEKS PERMINTAAN INFORMASI UNTUK APLIKASI ONLINE SHOP MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (STUDI KASUS: BENTO SHOP)

TUGAS AKHIR

Disusun untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada

Jurusan Sistem Informasi
Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember

Oleh:

DEA ANDIA RACHMAWATI

5212 100 177

Surabaya, May 2016

**KETUA
JURUSAN SISTEM INFORMASI**

Dr. Ir. Aris Fjanyanto, M.Kom.

NIP.19650310 199102 1 001



LEMBAR PERSETUJUAN

KLASIFIKASI TEKS PERMINTAAN INFORMASI UNTUK APLIKASI ONLINE SHOP MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (STUDI KASUS: BENTO SHOP)

TUGAS AKHIR

Disusun untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Jurusan Sistem Informasi
Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember

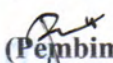
DEA ANDIA RACHMAWATI
5212 100 177

Disetujui Tim Penguji: Tanggal Ujian : May 2016
Periode Wisuda: September 2016

Renny Pradina K, S.T. , M.T


(Pembimbing 1)

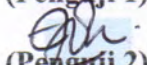
Radityo P.W, S.Kom. , M.Kom


(Pembimbing 2)

Nur Aini R., S.kom, M.Sc.Eng


(Penguji 1)

Irmasari Hafidz, S.Kom, M.Sc


(Penguji 2)

KLASIFIKASI TEKS PERMINTAAN INFORMASI UNTUK APLIKASI ONLINE SHOP MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE

Nama Mahasiswa : Dea Andia Rachmawati
NRP : 5212 100 177
Jurusan : SISTEM INFORMASI FTIF-ITS
Dosen Pembimbing 1 : Renny Pradina K, S.T. , M.T
Dosen Pembimbing 2 : Radityo P.W, S.Kom. , M.Kom

ABSTRAK

Pemanfaatan teknologi dalam bidang perdagangan dan penjualan diantaranya E-Commerce semakin berkembang. Berdasarkan data statistik dari ICD (lembaga penelitian dan informasi Media Group Digital) diketahui bahwa dari tahun 2012 – 2015 pasar E-commerce di indonesia meningkat sebanyak 42%. Salah satu pemanfaatan E-Commerce adalah forbento.com. Forbento merupakan semi E-commerce yang menjual bento tools (alat-alat untuk membuar bento) melalui website dan menghubungi customer service dengan menggunakan aplikasi pengiriman pesan singkat blackberry messenger.

Pada penelitian sebelumnya yang dilakukan oleh Hudalizaman mengenai pengembangan aplikasi personal assistant untuk membantu mengetahui informasi produk menggunakan pengolahan bahasa alami berbasis python (2015) telah dibuat aplikasi untuk menangani pertanyaan mengenai informasi produk yang diberikan dalam bentuk bahasa alami.

Namun pada aplikasi tersebut, belum dapat membedakan query dan non query secara otomatis karena sistem hanya mampu membaca kata yang termasuk ke dalam kategori produk saja. Jika harus mencari satu persatu tentunya akan

membutuhkan waktu yang lama. Oleh karena itu akan dilakukan klasifikasi teks sehingga dapat membedakan query (kata informasi produk) dan non query (non informasi produk) secara otomatis.

Metode yang digunakan dalam Tugas Akhir ini adalah Support Vector Machine. Pemilihan metode Support Vector Machine dalam klasifikasi teks ini dikarenakan metode Support Vector Machine (SVM) dapat memberikan solusi yang baik pada dataset yang besar dan meminimalisir terjadinya overfitting.

Tugas akhir ini menghasilkan model klasifikasi teks permintaan informasi yang memiliki nilai akurasi, presisi, recall dan F-Measure adalah 94.74%, 93.18%, 96.09%, dan 96.18%, sehingga hasil klasifikasi ini dapat dikategorikan baik. Dengan hasil ini diharapkan dapat membedakan teks query dan non query secara otomatis.

Kata kunci : E-Commerce, Forbento.com, Klasifikasi Teks, Informasi Produk, Support Vector Machine

TEXT CLASSIFICATION OF INFORMATION REQUEST FOR ONLINE SHOP USING SUPPORT VECTOR MACHINE ALGORITHM

Name : Dea Andia Rachmawati
NRP : 5212 100 177
Department : INFORMATION SYSTEM FTIF-ITS
Supervisor 1 : Renny Pradina K, S.T. , M.T
Supervisor 2 : Radityo P.W, S.Kom. , M.Kom

ABSTRACT

The use of technology in the field of trade and the sale of such E-Commerce is growing. Based on statistics from the ICD (research institutes and information Media Group Digital) note that from the year 2012-2015 E-commerce market in Indonesia increased by 42%. one of the utilization of E-Commerce is forbento.com. Forbento is a semi E-commerce that selling bento tools (tools for making bento) through the website and contact the customer service by using short messaging applications blackberry messenger.

In a previous study conducted by Hudalizaman regarding personal assistant application development to help determine product information using natural language processing based on python (2015) has made an application to handle questions regarding the product information provided in the form of natural language.

But in the application, can not distinguish between queries and non-queries automatically because the system is only able to read words that fall into the category of products only. If you should find one by one, it will certainly take a long time. Therefore, it will be the classification of text that can distinguish queries (words of information products) and non query (non-information product) automatically.

The method used in this final project is a Support Vector Machine. The selection method of Support Vector Machine in text classification is because Support Vector Machine (SVM) method can provide a good solution on large datasets and minimize overfitting.

The final task is to produce a text classification model requests for information that have value accuracy, precision, recall and F-Measure is: 94.74%, 93.18%, 96.09%, and 96.18%, so the results of this classification can be considered good. These results are expected to help distinguish non-text query and query automatically.

Keywords: E-Commerce, Forbento.com, Klasifikasi Teks, Informasi Produk, Support Vector Machine

KATA PENGANTAR

Puji syukur penulis panjatkan atas kehadiran Tuhan Yang Maha Esa atas segala berkat dan rahmat-Nya lah penulis dapat menyelesaikan buku tugas akhir dengan judul **“KLASIFIKASI TEKS PERMINTAAN INFORMASI UNTUK APLIKASI ONLINE SHOP MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*”** yang merupakan salah satu syarat kelulusan pada Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember Surabaya.

Secara khusus penulis akan menyampaikan ucapan terima kasih yang sedalam-dalamnya kepada:

1. Allah SWT yang telah memberi segala rahmat dan pencerahan untuk dapat menyelesaikan tugas belajar selama di Sistem Informasi ITS dan telah memberikan kemudahan serta kesehatan selama pengerjaan Tugas Akhir ini.
2. Kedua orang tua serta keluarga penulis yang selalu memberikan dukungan dan motivasi. Terima kasih atas doa dan dukungannya yang tiada henti.
3. Ibu Renny Pradina K, S.T., M.T dan Bapak Radityo P.W, S.Kom., M.Kom selaku dosen pembimbing yang telah meluangkan waktu dan pikiran di tengah kesibukan beliau untuk membimbing dan mengarahkan penulis dalam mengerjakan tugas akhir ini hingga selesai. Terima kasih atas waktu dan nasehatnya.
4. Ibu Nur Aini R.,S.kom, M.Sc.Eng dan Ibu Irmasari Hafidz, S.Kom, M.Sc selaku dosen penguji penulis yang selalu memberikan masukan yang meningkatkan kualitas dari Tugas Akhir ini.
5. Ibu Wiwik Anggraeni, S.Si., M.Kom selaku dosen wali penulis yang selalu memberikan motivasi dan saran selama penulis menempuh pendidikan S1.

6. Malik awab dan galih, yang sudah membantu penulis dan meyemangati penulis di masa masa sulit menjelang sidang.
7. Aditya Pramana yang selalu mendoakan, menyemangati penulis dan memberikan surprise yang tak terduga walaupun tidak bisa menemani penulis selama sidang berlangsung dan masa masa pengerjaan TA.
8. Para sahabat dekat yang selalu memberikan dukungan dan membantu penulis selama duduk dibangku perkuliahan sehingga bisa menyelesaikan Tugas Akhir ini (Widy, Nella, Desy, Janice, Danis, Piel).
9. Teman-teman dari RDIB, ADDI dan Solaris (SI-2012) yang menjadi rekan seperjuangan penulis dalam Tugas Akhir dan membantu penulis selama kuliah di Sistem Informasi.
10. Seluruh dosen pengajar, staff, dan karyawan di Jurusan Sistem Informasi, FTIF ITS Surabaya yang telah memberikan ilmu dan bantuan kepada penulis selama ini.
11. Serta semua pihak yang telah membantu dalam pengerjaan Tugas Akhir ini yang belum mampu penulis sebutkan diatas.

Terima kasih atas segala bantuan, dukungan, serta doanya. Semoga Tuhan senantiasa memberkati dan membalas kebaikan-kebaikan yang telah diberikan kepada penulis.

Penulis pun menyadari bahwa Tugas Akhir ini masih belum sempurna dengan segala kekurangan di dalamnya. Oleh karena itu penulis memohon maaf atas segala kekurangan yang ada di dalam Tugas Akhir ini dan bersedia menerima kritik dan saran. Semoga Tugas Akhir ini dapat bermanfaat bagi seluruh pembaca.

Surabaya, Juli 2016

DAFTAR ISI

ABSTRAK	iii
ABSTRACT	v
KATA PENGANTAR	vii
DAFTAR ISI	ix
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xv
DAFTAR KODE	xvii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan permasalahan	2
1.3 Batasan Permasalahan	3
1.4 Tujuan	3
1.5 Manfaat	3
1.6 Relevansi	3
BAB II TINJAUAN PUSTAKA	5
2.1 Studi Sebelumnya	5
2.2 Dasar Teori	10
2.2.1 Forbento.com	10
2.2.2 E-Commerce	11
2.2.3 Klasifikasi Teks	11
2.2.4 Support Vector Machine	14
2.2.5 <i>Kernel</i>	16
2.2.6 <i>Grid Search</i>	20
2.2.7 Evaluasi Performa Klasifikasi	20
BAB III METODE Pengerjaan Tugas Akhir	23
3.1 Penetapan Tujuan dan Studi Literatur	23
3.2 Pengumpulan Data	24
3.3 Tahap Praproses Teks	24
3.4 Tahap Klasifikasi	25
3.5 Evaluasi Hasil Uji	25
3.6 Analisa Hasil dan Pembahasan	25
3.7 Pembuatan Buku Tugas Akhir	25

BAB IV PERANCANGAN	27
4.1 Pengumpulan dan pre-processing data	27
4.1.1 Pengumpulan data.....	27
4.1.2 Pre-processing	28
4.2 Pembuatan Model Support Vector Machine.....	29
4.2.1 Menentukan Data Train dan Data Test.....	30
4.2.2 Membuat Dtm dan data frame	30
4.2.3 Membuat label query dan non query untuk Data Train dan Data Test	31
4.2.4 Penggunaan metode <i>grid search</i>	31
4.2.5 Membuat Model Klasifikasi SVM.....	32
BAB V IMPLEMENTASI	33
5.1 Implementasi Data.....	33
5.2 Proses Klasifikasi.....	34
5.2.1 Menginputkan Data	34
5.2.2 Praproses Teks.....	36
5.2.3 Menentukan Data Train dan Data Test.....	37
5.2.4 Pembuatan DTM dan Data Frame	37
5.2.5 Membuat Label Query dan Non Query	38
5.2.6 Klasifikasi menggunakan SVM.....	38
5.2.7 Penggunaan Metode <i>Grid Search</i>	39
5.2.8 Uji Model SVM.....	40
5.3 Word Frequency Distribution	41
BAB VI UJI COBA DAN ANALISIS HASIL	57
6.1 Membuat Model Uji Coba.....	57
6.1.1 Uji Coba I	58
6.1.2 Uji Coba III.....	60
6.1.3 Uji Coba IV	61
6.1.4 Uji Coba V.....	62
6.1.5 Uji Coba VI	62
6.2 Hasil Uji Coba Model.....	63
6.3 Uji Validasi.....	65
6.4 Analisis Hasil Uji Coba Model.....	67
6.4.1 Analisis Uji Validasi.....	67
6.4.2 Analisis Perbandingan Uji Coba.....	68
BAB VII KESIMPULAN DAN SARAN	87
7.1 Kesimpulan.....	87

7.2 Saran.....	88
DAFTAR PUSTAKA	89
BIODATA PENULIS	93
LAMPIRAN A (SKENARIO UJI COBA)	1
UJI COBA I.....	1
UJI COBA II	2
UJI COBA III.....	3
UJI COBA IV.....	4
UJI COBA V	6
UJI COBA VI.....	7
LAMPIRAN B (PENGUNAAN METODE GRID SEARCH).....	1
Kernel Linear	1
Kernel Linear 1	4
Kernel Radial 1	7
Kernel Radial	9
LAMPIRAN C (DAFTAR STOPWORDS)	1

Halaman ini sengaja dikosongkan

DAFTAR GAMBAR

Gambar 2.1 alur pemesanan melalui website.....	11
Gambar 2.2 gambar terbaik yang memisahkan Class -1 dan Class +1 (Romi Satria Wahono, 2015).....	15
Gambar 3.1 Diagram Alur Metodologi Penelitian	23
Gambar 4.1 Contoh Data Query dan Non Query	28
Gambar 4.2 Langkah pembuatan model Support Vector Machine.....	29
Gambar 4.3 Model Klasifikasi Support Vector Machine	32
Gambar 5.1 Grafik statistik data	35
Gambar 5.2 Grafik Distribusi Frekuensi Kata corpus lama ...	45
Gambar 5.3 Distribusi seluruh kata <i>corpus</i> lama	46
Gambar 5.4 Grafik Distribusi Frekuensi Kata <i>corpus</i> baru....	51
Gambar 5.5 Gambar distribusi seluruh kata <i>corpus</i> baru	52
Gambar 6.1 Grafik <i>KeUji</i> Coba II	59
Gambar 6.2 Grafik <i>Kernel Linear</i> 1	60
Gambar 6.3 Grafik <i>Kernel Radial</i> 1	61
Gambar 6.4 Grafik <i>Kernel Radial</i>	62

Halaman ini sengaja dikosongkan

DAFTAR TABEL

Tabel 2.1 Tabel Perbandingan Studi Sebelumnya.....	7
Tabel 2.2 Contoh <i>Case Folding</i>	12
Tabel 2.3 Contoh <i>Tokenizing</i>	13
Tabel 2.4 Contoh <i>Filtering</i>	14
Tabel 2.5 Fungsi Kernel pada SVM.....	16
Tabel 2.6 Confusion Matrix	21
Tabel 4.1 Contoh data sebelum dan setelah pre-processing...	29
Tabel 5.1 Hasil statistik.....	34
Tabel 5.2 Tabel hasil perubahan ke dalam huruf kecil.....	36
Tabel 5.3 Tabel hasil penghapusan angka dan tanda baca	37
Tabel 5.4 Top 50 Word Frequency Distribution <i>corpus</i> lama.....	41
Tabel 5.5 Pembakuan kata pada corpus	47
Tabel 5.6 Top 50 Word Frequency Distribution <i>corpus</i> lama.....	48
Tabel 5.7 50 Kata Teratas pada corpus lama.....	53
Tabel 5.8 50 Kata Teratas pada <i>corpus</i> baru	55
Tabel 6.1 Skenario Uji Coba	57
Tabel 6.2 Nilai akurasi berdasarkan parameter	63
Tabel 6.3 Nilai Akurasi <i>corpus</i> lama	64
Tabel 6.4 Nilai Akurasi <i>corpus</i> baru	64
Tabel 6.5 Tabel akurasi hasil uji coba model	65
Tabel 6.6 Tabel Hasil Uji Validasi Model	66
Tabel 6.7 Tabel perhitungan presisi, recall, F-Measure masing-masing kelas	68
Tabel 6.8 Perbandingan akurasi model uji coba II dan VI.....	68
Tabel 6.9 <i>Confusion Matrix</i> model II.....	69
Tabel 6.10 Teks yang diprediksi salah pada kelas query	69
Tabel 6.11 Teks yang diprediksi salah pada kelas non-query	71
Tabel 6.12 Confusion Matrik model Uji Coba V.....	74
Tabel 6.13 Teks yang diprediksikan salah pada kelas query	74
Tabel 6.14 Tabel perbandingan model lama dan baru	77
Tabel 6.15 Teks yang diprediksi salah pada kelas non-query	79
Tabel 6.16 Tabel perbandingan model lama dan baru	82

Halaman ini sengaja dikosongkan

DAFTAR KODE

Kode 5.1 Input data pada R.....	34
Kode 5.2 Merubah ke dalam bentuk corpus.....	36
Kode 5.3 Merubah ke dalam bentuk huruf kecil	36
Kode 5.4 Penghapusan angka dan tanda baca.....	36
Kode 5.5 Membagi data menjadi data train dan data test.....	37
Kode 5.6 Pembuatan dtm untuk data train dan data test	37
Kode 5.7 Pembuatan data frame	38
Kode 5.8 Pelabelan untuk data train dan data test.....	38
Kode 5.9 Klasifikasi dengan svm.....	39
Kode 5.10 Grid Search dengan kernel linear	39
Kode 5.11 Grid Search dengan Kernel Radial	40
Kode 5.12 Uji dan Mengukur model svm	41
Kode 5.13 50 Distribusi frekuensi kata	41
Kode 5.14 <i>Remove Stopwords</i> Bahasa Indonesia.....	44

Halaman ini sengaja dikosongkan

BAB I

PENDAHULUAN

Bab pendahuluan ini menjelaskan latar belakang masalah, rumusan masalah, batasan masalah, tujuan dan pengerjaan tugas akhir.

1.1 Latar Belakang

Perkembangan teknologi dan informasi pada era globalisasi saat ini semakin meningkat. Salah satunya pada bidang perdangan dan penjualan. Pemanfaatan teknologi dalam bidang perdagangan dan penjualan diantaranya adalah *E-Commerce*. Berdasarkan data statistik dari ICD (lembaga penelitian dan informasi Media Group Digital) diketahui bahwa dari tahun 2012 – 2015 pasar *E-commerce* di indonesia meningkat sebanyak 42% [1].

Pesatnya pertumbuhan E-Commerce di Indonesia didukung dengan data dari Kementrian Komunikasi dan Informasi diketahui bahwa nilai transaksi E-commerce pada tahun 2013 mencapai angka Rp 130 triliun. Pencapaian nilai transaksi yang tinggi tersebut berbanding lurus dengan jumlah pengguna internet di Indonesia yang mencapai angka 82 juta orang atau sekitar 30% dari total penduduk di Indonesia [1].

Dengan pesatnya perkembangan E-commerce tersebut membuat banyak orang yang mulai menggunakan E-commerce sebagai transaksi jual beli salah satunya adalah website *forbento.com*. *Forbento* merupakan semi E-commerce yang menjual *bento tools* (alat-alat untuk membuar *bento*) melalui website dan menghubungi *customer service* dengan meggunakan aplikasi pengiriman pesan singkat *blackberry messenger*.

Pada penelitian sebelumnya yang dilakukan oleh Hudalizaman mengenai pengembangan aplikasi *Personal Assistant* untuk membantu mengetahui informasi produk menggunakan pengolahan bahasa alami berbasis python

(2015) telah dibuat aplikasi untuk menangani pertanyaan mengenai informasi produk yang diberikan dalam bentuk bahasa alami.

Namun pada aplikasi tersebut, belum dapat membedakan query dan non query secara otomatis karena dalam sistem belum dilakukan klasifikasi sehingga hanya mampu membaca kata yang termasuk informasi produk. Jika pemilik harus mencari satu persatu teks yang termasuk ke dalam informasi produk tentunya akan membutuhkan waktu yang lama. Oleh karena itu akan dilakukan klasifikasi teks sehingga dapat membedakan query (kata yang menanyakan mengenai informasi produk) dan non query (kata yang tidak menanyakan mengenai informasi produk) secara otomatis.

Metode yang digunakan dalam Tugas Akhir ini adalah *Support Vector Machine*. Pemilihan metode *Support Vector Machine* dalam klasifikasi teks ini dikarenakan svm memiliki beberapa kelebihan yaitu, dapat memberikan solusi yang baik pada dataset yang besar dan meminimalisir terjadinya *overfitting*. *Overfitting* merupakan kemampuan model klasifikasi untuk melakukan klasifikasi data dengan sangat baik namun sangat buruk dalam melakukan klasifikasi data yang baru dan belum pernah ada. Dengan kelebihan tersebut maka SVM merupakan metode yang sesuai untuk mengklasifikasikan teks permintaan informasi produk [4]. Dengan dilakukannya klasifikasi teks diharapkan dapat membedakan *query dan non query* secara otomatis.

1.2 Rumusan permasalahan

Permasalahan yang dihadapi dalam penelitian ini antara lain adalah sebagai berikut:

1. Bagaimana cara melakukan praproses teks dan klasifikasi teks permintaan informasi produk?
2. Bagaimana hasil dan performa SVM dalam pengklasifikasian teks untuk permintaan informasi produk?

1.3 Batasan Permasalahan

Batasan dalam pengerjaan tugas akhir ini adalah :

1. Data yang digunakan berupa data *query* dan *non query* untuk klasifikasi permintaan informasi produk yang didapat dari website forbento.com.
2. Penelitian ini berfokus pada klasifikasi teks untuk permintaan informasi produk.
3. *Tools* yang digunakan adalah *package* e1071 pada program R- 3.2.2.
4. Output yang dihasilkan adalah model klasifikasi teks permintaan informasi produk.

1.4 Tujuan

Tujuan dari pengerjaan tugas akhir ini adalah :

1. Melakukan praproses teks dan klasifikasi teks permintaan informasi produk dengan menggunakan SVM.
2. Mengidentifikasi hasil dan performa SVM dalam klasifikasi teks permintaan informasi produk.

1.5 Manfaat

Manfaat dari pengerjaan tugas akhir ini adalah untuk membantu pemilik bisnis startup terutama mobile commerce untuk membedakan *query* dan *non query* secara otomatis sehingga tidak membutuhkan intervensi dari pemilik. Selain itu, tugas akhir ini bisa dijadikan sebagai masukan atau rujukan untuk penelitian-penelitian selanjutnya mengenai klasifikasi teks.

1.6 Relevansi

Relevansi pengerjaan tugas akhir ini terhadap area sistem informasi berada pada area Akuisisi Data dan Diseminasi Informasi dengan topik *Text Mining*. Area ini sesuai dengan penerapan beberapa matakuliah dari laboratorium terkait

seperti, Penggalan Data dan Analitika Bisnis, Sistem Cerdas, dan Sistem Pendukung Keputusan.

BAB II

TINJAUAN PUSTAKA

Bab ini berisi mengenai studi sebelumnya yang berhubungan dengan tugas akhir dan teori - teori yang berkaitan dengan permasalahan tugas akhir.

2.1 Studi Sebelumnya

Pada pengerjaan tugas akhir ini ada beberapa penelitian sebelumnya yang dijadikan acuan. Penelitian tersebut antara lain:

Pengembangan Aplikasi Personal Assistant Untuk Membantu Mengetahui Informasi Produk Menggunakan Pengolahan Bahasa Alami Berbasis Python oleh Hudalizaman (2015). Penelitian ini mengenai pembuatan aplikasi *Mobile Commerce* yang dapat menangani pertanyaan mengenai informasi produk yang diberikan dalam bentuk bahasa alami. Dibuatnya aplikasi ini karena pemilik merasa kesulitan dalam menangani permintaan informasi produk dari pelanggan. Hal ini terjadi karena selama ini setiap ada pembeli yang menanyakan mengenai informasi produk tertentu, penjual harus terlebih dahulu membuka toko online yang dimiliki, *login* sebagai pemilik, dan kemudian mengecek informasi produk yang dimiliki. Tentunya ini menyebabkan transaksi jual-beli yang terjadi menjadi terhambat.

Cara kerja sistem adalah, chat dari pembeli yang menanyakan informasi produk, akan penjual masukkan ke dalam sistem, dan sistem akan secara otomatis memberikan balasan berupa informasi produk dalam bentuk email. Dalam menjawab pertanyaan mengenai produk, dilakukan klasifikasi bahasa alami dengan mengkategorisasikan kalimat pesan menjadi kalimat *query* (informasi produk) dan *non query* (*non* informasi produk). Kata yang termasuk ke dalam kata *non query* akan dihapuskan sehingga dalam sistem hanya akan terdapat kata yang termasuk ke dalam kata *query*.

Namun dalam aplikasi ini masih terdapat kekurangan yaitu, penjual harus mencari teks pelanggan yang menanyakan

informasi produk. Hal ini terjadi karena teknis untuk mendapatkan jawaban informasi produk adalah dengan memasukkan chat pelanggan yang menanyakan informasi produk ke dalam sistem. Jika hanya sedikit pesan teks yang masuk, tentunya akan lebih mudah untuk mencari teks pelanggan yang menanyakan informasi produk, namun jika jumlah pesan teks yang masuk sangatlah banyak tentunya akan menyulitkan penjual jika harus mencari satu-persatu. Pencarian ini tentunya akan membutuhkan waktu yang lama dan dapat menghambat transaksi jual beli.

Dari hasil pembuatan aplikasi terdapat 3 uji coba skenario, yaitu skenario pertama untuk mengetahui performa aplikasi dalam menjawab pertanyaan pelanggan terkait produk yang diharapkan, skenario kedua untuk mengetahui performa ketika produk ada pada kalimat atau tidak, dan yang terakhir untuk mengetahui kecepatan sistem dalam menangani *request*. Hasil uji skenario pertama diketahui nilai *recall* adalah 81%, *accuracy* adalah 89% dan *precision* adalah 67%. Hasil uji skenario kedua diketahui nilai rata-rata dari *recall* adalah 64%, *accuracy* adalah 71.31% dan *precision* adalah 61%. Hasil uji kecepatan sistem diketahui bahwa dalam memproses setiap email yang masuk membutuhkan waktu 3.44 detik.

Indonesian News Classification Using Support Vector Machine oleh Dewi Y.Lilian, Agung Hardianto, M.Ridok (2011). Penelitian ini mengenai klasifikasi berita bahasa indonesia ke dalam 4 kategori yaitu, national, international, business and finance, dan sports menggunakan kernel Radial Basis Function (RBF). Penelitian ini mencari parameter C (Complexity) dan γ terbaik untuk menghasilkan akurasi SVM terbaik. Dari hasil penelitian didapat tingkat rata-rata akurasi SVM adalah 85%, dengan nilai $C = 110$ dan $\gamma = 1$.

Klasifikasi Dokumen Berita Menggunakan Metode Support Vector Machine Dengan Kernel Radial Basis Function oleh Adyatma Bhaskara Hutomo (2014). Penelitian ini mengenai klasifikasi dokumen berita berbahasa inggris ke dalam 2 kelas yaitu kelas earn dan kelas -earn. Pada penelitian ini dilakukan pemilihan fitur chi untuk menentukan kata yang cocok untuk

dijadikan penciri dalam pembuatan model klasifikasi dan kemudian dilakukan pencarian parameter terbaik untuk kernel RBF, yaitu nilai C (Complexity) dan setelah itu dilakukan perbandingan pembobotan menggunakan metode *tf*(*term frequency*) dan *tf-idf*(*term frequency – inverse document frequency*). Dari hasil perbandingan didapat Hasil akurasi dengan menggunakan pembobotan *tf-idf* (*term frequency – inverse document frequency*) sebesar 92.97% sedangkan hasil akurasi dengan menggunakan pembobotan *tf* (*term frequency*) sebesar 93.21%.

Klasifikasi Kondisi Penderita Penyakit Hepatitis Dengan Menggunakan Metode Support Vector Machine oleh Lailil Muflikha, Achmad Ridok, Jendi Hardono (2013). Penelitian ini mencari nilai C (*Complexity*), yaitu parameter yang digunakan untuk mengukur tingkat akurasi klasifikasi dari metode SVM, dengan menggunakan 2 atribut yaitu klasifikasi 19 atribut dan 15 atribut. Hasil rata-rata akurasi menggunakan dataset Hepatitis dengan 19 atribut dengan nilai C = 30 adalah 82.08%. Hasil rata-rata akurasi menggunakan dataset Hepatitis dengan 15 atribut dengan nilai C = 40, 50, 60 adalah 84.93%.

Tabel 2.1 Tabel Perbandingan Studi Sebelumnya

Judul	Penulis	Tujuan	Hasil Penelitian
Pengembangan Aplikasi Personal Assistant Untuk Membantu Mengetahui Informasi Produk Menggunakan	Hudalizaman (2015)	Membuat aplikasi personal assistant untuk menangani pertanyaan mengenai informasi produk yang dalam bentuk bahasa alami.	Hasil uji skenario pertama diketahui nilai <i>recall</i> adalah 81%, <i>accuracy</i> adalah 89% dan <i>precision</i> adalah 67%.

Judul	Penulis	Tujuan	Hasil Penelitian
Pengolahan Bahasa Alami Berbasis Python			Hasil uji skenario kedua diketahui nilai rata-rata dari <i>recall</i> adalah 64% , <i>accuracy</i> adalah 71.31% dan <i>precision</i> adalah 61%. Hasil uji kecepatan sistem diketahui bahwa dalam memproses setiap email yang masuk membutuhkan waktu 3.44 detik.
Indonesian News Classification Using Support Vector Machine	Dewi Y.Lilian, Agung Hardianto, M.Ridok (2011)	Mengklasifikasi kan Artikel Berita Berbahasa Indonesia ke dalam 4 kategori menggunakan Support Vector Machine dengan kernel Radial Basis Function	Tingkat rata-rata akurasi dengan menggunakan SVM adalah 85% dan Nilai parameter terbaik yang

Judul	Penulis	Tujuan	Hasil Penelitian
		(RBF).	menghasilk an tingkat rata-rata akurasi tertinggi adalah C (<i>Complexity</i>) = 110 dan gamma SVM = 1.
Klasifikasi Dokumen Berita Menggunaka n Metode Support Vector Machine Dengan Kernel Radial Basis Function	Adyatma Bhaskara Hutomo (2014)	Pengklasifikasia n Dokumen berita berbahasa inggris menggunakan metode kernel <i>radial basis function</i> , menggunakan pemilihan fitur Chi, dan membandingkan <i>tf-idf</i> , <i>tf</i> sebagai metode pembobotan.	Hasil akurasi dengan menggunak an pembobotan <i>tf-idf</i> (<i>term frequency – inverse document frequency</i>) sebesar 92.97% sedangkan hasil akurasi dengan menggunak an pembobotan <i>tf</i> (<i>term frequency</i>) sebesar 93.21%
Klasifikasi	Lailil	Pengklasifikasia	Hasil rata-

Judul	Penulis	Tujuan	Hasil Penelitian
Kondisi Penderita Penyakit Hepatitis Dengan Menggunakan Metode Support Vector Machine	Muflikha, Achmad Ridok, Jendi Hardono (2013)	n Penyakit Hepatitis dengan mencari nilai C (<i>Complexity</i>) terbaik yang dibagi ke dalam 2 atribut berbeda yaitu, klasifikasi 19 atribut dan kalsifikasi 15 atribut.	rata akurasi menggunakan dataset Hepatitis dengan 19 atribut dengan nilai $C = 30$ adalah 82.08%. Hasil rata-rata akurasi menggunakan dataset Hepatitis dengan 15 atribut dengan nilai $C = 40, 50, 60$ adalah 84.93%.

2.2 Dasar Teori

2.2.1 Forbento.com

Forbento didirikan sejak januari 2011 oleh Rahayu Fatmawati , yang merupakan perusahaan yang menjual *bento tools*, yaitu peralatan yang dapat mempermudah dalam membuat bekal dan buku cara membuat *bento*. Forbento tidak hanya menjual alat alat *bento* namun juga memberikan tips-tips mengenai cara pembuatan *bento*, penggunaan alat *bento tools*, dan tips dalam menjalankan bisnis *catering bento*.

2.2.2 E-Commerce

E-commerce adalah suatu cara berbelanja atau berdagang secara online yang memanfaatkan fasilitas internet dimana terdapat website yang menyediakan layanan get and deliver [2][3]. Forbento merupakan semi e-commerce yang melakukan transaksi jual beli dengan 2 cara, yaitu dengan pembelian langsung dari websitenya dan dengan menghubungi *customer service* menggunakan pesan singkat pada aplikasi *blackberry messenger*.

Alur kerja pemesanan melalui website :



Gambar 2.1 alur pemesanan melalui website

Pada Tugas akhir ini data yang digunakan merupakan data yang didapatkan dari *customer service* berupa pesan teks singkat dan dari website *forbento.com* yang akan diolah untuk dilakukan klasifikasi.

2.2.3 Klasifikasi Teks

Klasifikasi teks merupakan bagian penting dari text mining yang termasuk ke dalam pembelajaran jenis *supervised learning* [6]. Klasifikasi teks adalah sebuah proses untuk mengkategorisasikan sebuah teks sesuai dengan kategori yang telah ditentukan yang bertujuan untuk mempermudah dalam mengorganisir teks dalam jumlah besar [7].

Dalam text mining, klasifikasi mengacu kepada aktifitas menganalisis atau mempelajari himpunan dokumen teks

untuk memperoleh suatu model atau fungsi yang dapat digunakan untuk mengelompokkan dokumen teks lain yang belum diketahui kelasnya ke dalam satu atau lebih kelas [6] [8]. Data yang digunakan dalam klasifikasi teks terdiri dari 2 data, yaitu data training dan data testing. Data training digunakan untuk membangun model atau fungsi sedangkan data testing digunakan untuk mengetahui keakuratan model atau fungsi yang akan dibangun pada data training. [6]

a) Praproses teks

Praproses teks merupakan tahap yang dilakukan sebelum melakukan proses pengelompokan dokumen. Pada tahap praproses ini dilakukan beberapa subproses agar dokumen dapat dipakai untuk melakukan proses pengelompokan. Tujuan dari praproses teks adalah untuk menyeragamkan bentuk kata dan mengurangi volume kosakata. Tahap ini terdiri dari : [7]

- *Case Folding* yaitu proses dalam mengubah semua huruf dalam teks menjadi huruf kecil. Karakter selain huruf akan dihilangkan.

Tabel 2.2 Contoh *Case Folding*

Teks Input	Teks Output
Di produk mba, pilihannya ada apa aja utk rice mold dan bento cutter?	di produk mba pilihannya ada apa aja utk rice mold dan bento cutter

- *Tokenizing* yaitu sebuah proses untuk memisahkan setiap kata dalam suatu kalimat sehingga menghasilkan kumpulan kata-kata yang

berdiri sendiri. Pemisahan kata dilakukan dengan cara menemukan spasi (space) antar kata.

Tabel 2.3 Contoh *Tokenizing*

Teks Input	Teks Output
di produk mba pilihannya ada apa aja utk rice mold dan bento cutter	di produk mba pilihannya ada apa aja utk rice mold dan bento cutter

- *Filtering* yaitu proses untuk mengambil kata penting dari hasil token. Dalam melakukan filtering dapat menggunakan stoplist atau wordlist (menyimpan kata penting). *Stoplist / stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah “yang”, “dan”, “di”, “dari” dan seterusnya. Dengan menggunakan daftar *stoplist*, maka setiap kata dalam koleksi akan dicocokkan dengan kata-kata

yang ada dalam *stoplist*. Apabila terdapat kata yang sama, maka kata itu akan dibuang dari koleksi.

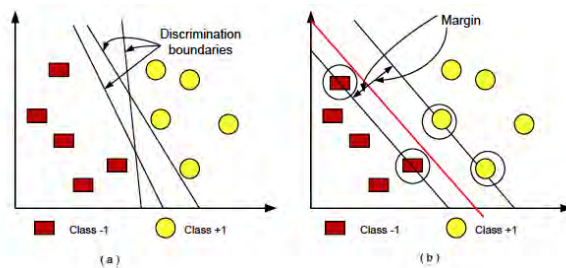
Tabel 2.4 Contoh *Filtering*

Teks Input	Teks Output
Di	produk
produk	pilihannya
mba	rice
pilihannya	mold
ada	bento
apa	cutter
aja	
utk	
rice	
mold	
dan	
bento	
cutter	

2.2.4 Support Vector Machine

Support vector machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory [4]. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada *input space*. Gambar 1.a memperlihatkan beberapa pattern yang merupakan

anggota dari dua buah class : $+1$ dan -1 . Pattern yang tergabung pada *class* -1 disimbolkan dengan warna merah (kotak), sedangkan pattern pada *class* $+1$, disimbolkan dengan warna kuning (lingkaran). Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. Garis pemisah (*discrimination boundaries*) ditunjukkan pada Gambar 1.a merupakan salah satu alternatif garis pemisah yang memisahkan kedua *class* [4] [10] [11] [12].



Gambar 2.2 gambar terbaik yang memisahkan Class -1 dan Class +1 (Romi Satria Wahono, 2015)

Hyperplane pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur margin hyperplane tersebut dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing *class*. Pattern yang paling dekat ini disebut sebagai support vektor. Garis solid pada Gambar 1.b menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua *class*, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah support vector. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pembelajaran pada support vector machine [4] [9] [10] [11].

2.2.5 Kernel

Beberapa metode dalam analisis data mining banyak menggunakan fungsi *linear*. Namun masalah dalam dunia nyata jarang yang bersifat *linear* kebanyakan bersifat non *linear*. Sehingga untuk mengatasinya dengan cara mentransformasikan data ke dalam dimensi ruang yang lebih tinggi. SVM dapat digunakan pada data non *linear* dengan menggunakan *Kernel Trick* [12].

Konsep dari *kernel trick* adalah memetakan data yang bersifat non-linear pada input space ke ruang vektor baru yang berdimensi lebih tinggi dimana kedua class dapat dipisahkan secara linear oleh sebuah *hyperplane*. *Kernel Trick* dirumuskan sebagai berikut : [12]

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)$$

Keterangan :

K = menunjukkan fungsi kernel

\vec{x}_i = menunjukkan vektor data latih

\vec{x}_j = menunjukkan vektor data uji

$\Phi(.)$ = fungsi pemetaan dari ruang input ke ruang fitur.

Dalam SVM terdapat beberapa Fungsi *Kernel* yang biasa dipakai , yaitu : [12].

Tabel 2.5 Fungsi Kernel pada SVM

Kernel	Penjelasan
<i>Linear</i>	<p>Rumus : $x^T x$, dengan x adalah <i>data training</i>.</p> <p>Ciri : [13]</p> <ul style="list-style-type: none"> ▪ Cocok untuk klasifikasi teks karena kebanyakan teks

Kernel	Penjelasan
	<p>terpisah secara linear</p> <ul style="list-style-type: none"> ▪ Cocok digunakan jika jumlah fitur besar <p>Contoh menggunakan Dataset Iris :</p> <pre> Parameters: SVM-Type: C-classification SVM-Kernel: linear cost: 1 gamma: 0.25 Number of Support Vectors: 29 </pre>
<i>Polynomial</i>	<p>Rumus : $(x^T x_i + 1)^p$, dengan x dan x_i adalah pasangan dua data training, p konstanta dengan nilai > 0</p> <p>Ciri : [14]</p> <ul style="list-style-type: none"> ▪ Dapat memperluas fitur tanpa meningkatkan biaya komputasi ▪ Tidak memberikan tingkat akurasi yang tinggi dalam training atau testing <p>Contoh menggunakan Dataset Iris :</p>

Kernel	Penjelasan
	<p>Parameters:</p> <p>SVM-Type: C-classification</p> <p>SVM-Kernel: polynomial</p> <p>cost: 1</p> <p>degree: 3</p> <p>gamma: 0.25</p> <p>coef.0: 0</p> <p>Number of Support Vectors: 54</p>
<p><i>Radial Basis Function</i> (RBF)</p>	<p>Rumus : $\exp(-\frac{1}{2\sigma^2} \ x - x_i\ ^2)$, dengan x dan x_i adalah pasangan dua data training, σ adalah konstanta dengan nilai > 0</p> <p>Ciri : [15]</p> <ul style="list-style-type: none"> ▪ Memiliki performa yang bagus dalam melakukan klasifikasi ▪ Memiliki rentang nilai kecil ▪ Memiliki perilaku seperti kernel sigmoid dengan parameter tertentu <p>Contoh menggunakan dataset Iris :</p>

Kernel	Penjelasan
	<p>Parameters:</p> <p>SVM-Type: C-classification</p> <p>SVM-Kernel: radial</p> <p>cost: 1</p> <p>gamma: 0.25</p> <p>Number of Support Vectors: 51</p>
<p><i>Tangent</i> <i>Hyperbolic</i> (Sigmoid)</p>	<p>Rumus : $\tanh(\beta x^T x_i + \beta_1)$, dimana $\beta, \beta_1 \in R$, dengan x dan x_i adalah pasangan dua data training</p> <p>Ciri : [16]</p> <ul style="list-style-type: none"> • Kerja mirip dengan kernel RBF dalam parameter tertentu • Sulit untuk menentukan parameter yang cocok untuk kernel ini <p>Contoh menggunakan Dataset Iris :</p> <p>Parameters:</p> <p>SVM-Type: C-classification</p> <p>SVM-Kernel: sigmoid</p> <p>cost: 1</p> <p>gamma: 0.25</p> <p>coef.0: 0</p> <p>Number of Support Vectors: 54</p>

Berdasarkan tabel 2 maka metode kernel yang akan digunakan dalam Tugas Akhir ini adalah kernel *Linear* dan *Radial Basis Function* (RBF).

2.2.6 Grid Search

Grid Search merupakan salah satu algoritma yang sering digunakan untuk estimasi parameter. Prinsip *kerja Grid Search* adalah menentukan beberapa nilai pada rentang tertentu dan melakukan pencarian pada rentang tersebut sampai didapatkan hasil yang optimal. *Grid Search* bertujuan untuk membuat grid parameter dari setiap pasangan C, γ (cost dalam pembentukan model (tahap pelatihan / *training*), dan γ merupakan parameter yang digunakan untuk kernel). Pasangan nilai parameter yang terbaik dapat diukur dengan menggunakan *cross-validation*. *Cross-Validation* adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara *random* ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan. Parameter yang sudah optimal dapat digunakan sebagai model SVM terbaik. [17] [18]

2.2.7 Evaluasi Performa Klasifikasi

Pengujian akan dilakukan pada hasil klasifikasi menggunakan SVM untuk mengetahui akurasi klasifikasi SVM terhadap suatu data uji. Pengukuran tersebut didapatkan dalam sebuah set *confusion matrix*. *Confusion matrix* merupakan sebuah tabel yang terdiri dari banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi yang digunakan untuk menentukan kinerja model klasifikasi. [12] [19]

Tabel 2.6 Confusion Matrix

		Kelas Prediksi	
		Positif	Negatif
Observasi	Positif	TP	FN
	Negatif	FP	TN

Keterangan :

- *TP (True Positive)* adalah kelas yang diprediksi positif dan benar.
- *TN (True Negatif)* adalah kelas yang diprediksi negatif dan benar.
- *FP (False Positive)* adalah kelas yang diprediksi positif dan salah.
- *FN (False Negatif)* adalah kelas yang diprediksi negative dan salah.

Dari *confusion* tabel di atas kemudian dapat juga diukur tingkat akurasi, presisi, dan recall.

a) Presisi

Presisi digunakan untuk mengetahui banyaknya item yang dikategorikan ke dalam kategori yang seharusnya. Presisi dapat dihitung dengan menggunakan rumus berikut : [20]

$$\text{Presisi} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

b) Recall

Recall digunakan untuk mengetahui banyaknya item yang diprediksikan benar. *Recall* dapat dihitung dengan menggunakan rumus berikut : [20]

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

c) Akurasi

Akurasi dari klasifikasi digunakan untuk melihat kinerja secara keseluruhan. Akurasi dapat dihitung dengan menggunakan rumus berikut :
[19]

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

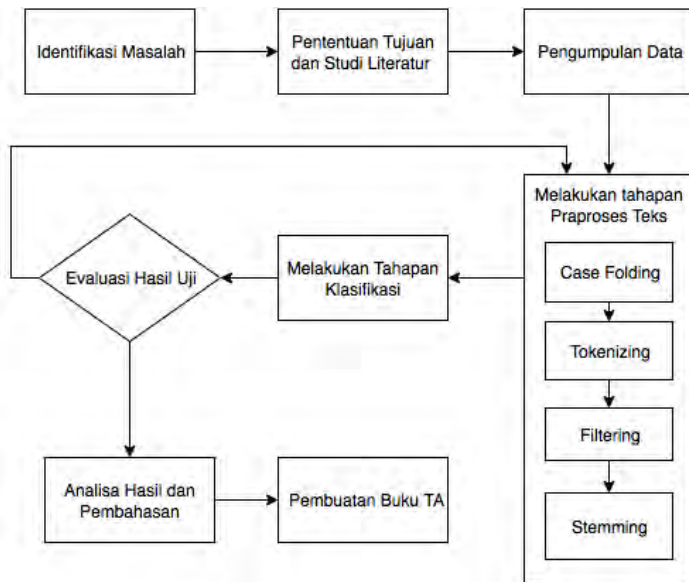
Selain itu digunakan pengukuran menggunakan F-Measure. F-Measure merupakan pengukuran yang mengkombinasikan presisi dan recall yang digunakan untuk mengukur keberhasilan information retrieval. parameter untuk F-Measure dapat dihitung dengan menggunakan rumus berikut : [20]

$$F - Measure = \frac{2 \cdot Recall \cdot Presisi}{Recall + Presisi}$$

BAB III

METODE Pengerjaan Tugas Akhir

Pada bab ini akan dijelaskan mengenai langkah-langkah sistematis yang dilakukan dalam tugas akhir agar terlaksana dengan terstruktur. Diagram alir metodologi tugas akhir dapat dilihat pada Gambar 3.1:



Gambar 3.1 Diagram Alur Metodologi Penelitian

3.1 Penetapan Tujuan dan Studi Literatur

Pada tahapan ini dilakukan penentuan tujuan, yaitu, menentukan tujuan dan batasan masalah dari penelitian tugas akhir serta mencari tinjauan pustaka mengenai konsep klasifikasi teks menggunakan metode *support vector machine* yang digunakan untuk menyelesaikan permasalahan pada tugas akhir ini.

3.2 Pengumpulan Data

Pada tahapan ini dilakukan pengumpulan data-data yang dibutuhkan untuk pengerjaan Tugas Akhir yang dilakukan. Data yang digunakan merupakan data *query* dan *non query* untuk permintaan informasi produk yang didapat dari website forbento.com. Data Query merupakan data Informasi Produk dan Data Non Query merupakan data Bukan Informasi Produk. Data informasi Produk merupakan data mengenai pertanyaan pelanggan terkait produk yang dijual sedangkan data yang bukan informasi produk merupakan data yang tidak menanyakan mengenai informasi produk. Data ini kemudian diolah sehingga dapat digunakan untuk proses klasifikasi teks. Contoh Data yang termasuk Data Informasi Produk dan Bukan Informasi Produk dapat dilihat pada Tabel 3.1.

Table 3.1 Contoh Data Informasi dan Bukan Informasi Produk

Data Informasi Produk	Data Bukan Informasi Produk
Di produk mba,pilihannya ada apa aja utk rice mold dan bento cutter?	Eh murah ya
Drawing food kpn ada mbak?	Wah good idea tuh
Sis jual citakannya telur puyu?	Mb Ayu..salam kenal yaa

3.3 Tahap Praproses Teks

Dalam penelitian ini pengolahan bahasa dilakukan dengan praproses teks untuk menyeragamkan bentuk kata dan mengurangi volume kosakata. Tahap ini terdiri dari :

- *Case Folding* yaitu proses dalam mengubah semua huruf dalam teks menjadi huruf kecil. Karakter selain huruf akan dihilangkan.

- *Tokenizing* yaitu sebuah proses untuk memisahkan setiap kata dalam suatu kalimat sehingga menghasilkan kumpulan kata-kata yang berdiri sendiri.
- *Filtering* yaitu proses untuk mengambil kata penting dari hasil token. Dalam melakukan filtering dapat menggunakan *stoplist* atau *wordlist* (menyimpan kata penting). *Stoplist / stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*.

3.4 Tahap Klasifikasi

Pada tahap ini dilakukan klasifikasi pada permintaan informasi produk menggunakan metode Support Vector Machine (SVM) menggunakan tools R 3.2.2 dan library yang digunakan adalah e1071. Klasifikasi teks yang dilakukan akan dibagi menjadi 2 kategori yaitu, kategori *query* dan *non query* dengan hasil berupa model klasifikasi. Data yang digunakan dalam klasifikasi akan dibagi menjadi 2, yaitu data *training* dan data *testing* dengan perbandingan 70 : 30.

3.5 Evaluasi Hasil Uji

Pada tahap ini dilakukan evaluasi hasil uji untuk mengetahui performa SVM dalam melakukan klasifikasi. Untuk menghitung performa SVM dilihat dari akurasi, presisi, *recall* dan *F-measure*.

3.6 Analisa Hasil dan Pembahasan

Pada tahap ini dilakukan analisa dari Hasil Uji proses klasifikasi teks permintaan informasi produk yang telah dilakukan sebelumnya dan akan dibuat pembahasan mengenai hasil tersebut.

3.7 Pembuatan Buku Tugas Akhir

Tahap ini merupakan tahap akhir dari penelitian. Pada tahap ini dilakukan dokumentasi untuk penulisan laporan tugas akhir. Hasil dari laporan tugas akhir berupa buku yang berisi

keseluruhan proses yang dilakukan dalam penelitian ini. Tugas Akhir ini diharapkan bisa dijadikan sebagai rujukan penelitian berikutnya mengenai klasifikasi teks.

BAB IV PERANCANGAN

Bab ini menjelaskan tentang rancangan penelitian tugas akhir untuk membuat model klasifikasi. Bab ini berisikan proses pengumpulan data, pengolahan data, dan perancangan model.

4.1 Pengumpulan dan pre-processing data

Pada subbab ini dilakukan pengumpulan dan pre-processing data. Pengumpulan data merupakan data yang digunakan untuk tugas akhir ini, dan pre-processing merupakan tahap yang dilakukan untuk mengolah data sebelum digunakan untuk pembuatan model menggunakan svm.

4.1.1 Pengumpulan data

Pengumpulan data yang digunakan pada tugas akhir ini bersumber dari Website Forbento.com. Data berupa data text pesan singkat dari aplikasi *blackberry messenger*. Oleh Hudalizaman (2011), data yang diambil berupa data percakapan yang dilakukan oleh pembeli sehingga didapatkan data dengan isi :

- msg_id
- file_id
- text, dan
- numchars

Kemudian data tersebut diberi label untuk dikelompokkan menjadi kelompok query dan non query. Berdasarkan penjelasan di atas, maka didapat data dengan isi :

- msg_id,
- file_id,
- text,
- numchars,

- label.

Contoh hasil data yang didapatkan penulis dapat dilihat pada Gambar 4.1 :

msg_id	file_id	text	numchars	query
id201304041	209EA856.cs	Mau nanya hrg ring pancake telur hello kity brp ya	54	query
id201304041	209EA856.cs	Kalo rice mold hello kity 3 set brp ya say	42	query
id201304041	209EA856.cs	Kalo vegetable cutter kena brp say	34	query
id201212091	20D4C2F2.cs	Buku ibento edisi 2 bs pesan?	29	query
id201302031	20D4C2F2.cs	Ada contohnya bento sleeping bear?:). Thx	41	query
id201303111	20D4C2F2.cs	Sis.. Itu cetakan micky brp?	28	query
id201303111	20D4C2F2.cs	Punya cetakan apa lg? Kyk jepretan mata,dll	43	query
id201303111	20D4C2F2.cs	Bisa tlg minta foto2 cetakan lain?	34	query
id201303121	20D4C2F2.cs	Cetakan nasi mksdnya	20	query
id201303121	20D4C2F2.cs	Klo cutter yg bentuk2 bunga ato bentuk laen ada?	48	query
id201303121	20D4C2F2.cs	Maap..coba sy mau liat cutter2..	32	query
id201303131	20D4C2F2.cs	Sis, maap..sy mau liat foto cutter2..	37	query
id201303131	20D4C2F2.cs	Cutter bentuk bunga, tema garden gitu	38	query
id201303131	20D4C2F2.cs	Klo eadible pen ada?	20	query
id201303131	20D4C2F2.cs	Cutter varous char kyk gmn?	27	query
id201303131	20D4C2F2.cs	Yg tipe2 kyk flower?	20	query
id201302261	20F5CE60.cs	Ricemold isi 3+cutternya mba?	29	nonquery
id201303221	20F7B0AF.cs	selamat siang mbak..	20	nonquery
id201303221	20F7B0AF.cs	ke salatiga jawa tengah	23	nonquery
id201303221	20F7B0AF.cs	Ongkir brapa mbak ke salatiga?	30	nonquery
id201303221	20F7B0AF.cs	Ok	2	nonquery
id201303221	20F7B0AF.cs	Sbntar saya pilih produknya :)	30	nonquery
id201303221	20F7B0AF.cs	Minta nomer ac bca ya :)	24	nonquery
id201303221	20F7B0AF.cs	Ohya pengiriman brapa hari ya mbak?	35	nonquery
id201303221	20F7B0AF.cs	ok2..	5	nonquery
id201303221	20F7B0AF.cs	Ada diskon ndak mbak? Xixixix	30	nonquery
id201303221	20F7B0AF.cs	Nanti kalo sdh transfer tak info ya mbak :)	43	nonquery
id201303221	20F7B0AF.cs	Xixixixi	9	nonquery
id201303221	20F7B0AF.cs	ok2	3	nonquery
id201303221	20F7B0AF.cs	Ndak pa2 mbak.. :)	18	nonquery
id201303221	20F7B0AF.cs	Nanti saya info kalau sdh transfer ya.m	39	nonquery
id201303221	20F7B0AF.cs	Selamat malam.. Mbak..	22	nonquery
id201303221	20F7B0AF.cs	Sudah saya kirim pembayaranya 45rbu ke rekenin	51	nonquery

Gambar 4.1 Contoh Data Query dan Non Query

Pada Tugas Akhir ini variabel yang akan digunakan adalah variabel *text* dan *label* dengan jumlah data sebanyak 9680 *dataset*.

4.1.2 Pre-processing

Setelah dilakukan pengumpulan data kemudian dilakukan praproses teks. Tahapan ini dilakukan untuk menyeragamkan

bentuk kata dan mengurangi volume kosakata agar dapat dipakai untuk melakukan proses pengelompokkan [7]. Praproses teks akan menggunakan aplikasi R Studio.

Dalam tahap praproses semua data teks akan diubah menjadi :

- Bentuk huruf kecil
- Dilakukan penghapusan angka
- Dan dilakukan penghapusan tanda baca

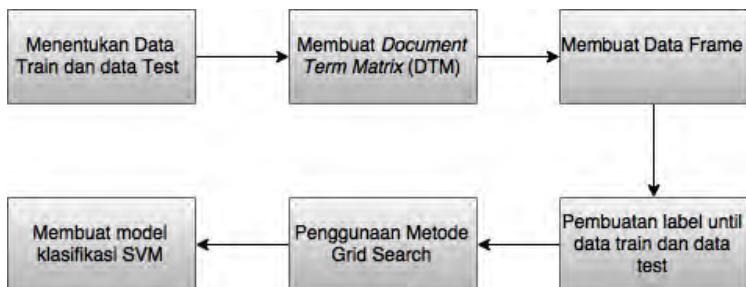
Contoh data sebelum di praproses dan setelah di praproses dapat dilihat pada Tabel 4.1:

Tabel 4.1 Contoh data sebelum dan setelah pre-processing

Sebelum <i>pre-processing</i>	Setelah <i>pre-processing</i>
Sis...kalo buku bonitanya itu PO atau ready stock?	siskalo buku bonitanya itu po atau ready stock
Bonita harga brp sis?	bonita harga brp sis
Ibu , saya bisa minta contoh bento box untuk wedding	ibu saya bisa minta contoh bento box untuk wedding

4.2 Pembuatan Model Support Vector Machine

Dalam pembuatan model support Vector Machine langkah yang dilakukan dapat dilihat pada Gambar 4.2 [21]



Gambar 4.2 Langkah pembuatan model Support Vector Machine

Berdasarkan Gambar 4.2 diketahui bahwa dalam pembuatan model *Support Vector Machine* terdapat 6 tahapan, yaitu :

- Menentukan Data Train dan Data Test
- Membuat dtm dan data frame
- Membuat label untuk data train dan data test
- Mencari nilai C terbaik untuk kernel liner dan mencari pasangan C dan γ terbaik untuk kernel radial (penggunaan metode *grid search*).
- Membuat model klasifikasi SVM

Untuk masing-masing penjelasan dari langkah di atas akan dijelaskan pada subab dibawah.

4.2.1 Menentukan Data Train dan Data Test

Dalam pengolahan data dibutuhkan dua set data, yaitu data *train* dan data *test*. Data *train* digunakan untuk membuat model klasifikasi, sedangkan data *test* digunakan untuk menguji akurasi model yang didapatkan [22]. Pada Tugas Akhir ini data dibagi menjadi data *train* dan data *test* dengan perbandingan 70 : 30.

4.2.2 Membuat Dtm dan data frame

Data yang sudah dibagi ke dalam bentuk data train dan data test akan dirubah ke dalam bentuk *document term matrix* (dtm). Dtm merupakan sebuah matriks dokumen yang mewakili hubungan antara kata dan dokumen, di mana setiap baris merepresentasikan kata dan setiap kolom untuk dokumen. Pada dtm *corpus* dalam bentuk teks akan diubah menjadi objek matematik yang dapat dianalisis menggunakan teknik kuantitatif. Contohnya, untuk mendapatkan frekuensi kata yang paling sering muncul dalam *corpus*. [23]

Kemudian untuk masing-masing dtm akan dirubah ke dalam bentuk data frame untuk bisa digunakan dalam klasifikasi menggunakan svm. Data frame digunakan untuk

menyimpan data dalam bentuk tabel. data yang disimpan dalam data frame berupa data vektor, sehingga lebih mudah untuk mengatur data dan menerapkan fungsi ke pada data frame. Dalam Tugas Akhir ini, data frame akan digunakan sebagai variabel untuk pembuatan model svm. [24]

4.2.3 Membuat label query dan non query untuk Data Train dan Data Test

Pada tahapan ini data label yang ada pada data awal dibagi ke dalam 2 label, yaitu label *train* dan label *test*. Hal ini dilakukan karena pada data text yang merupakan variabel target yang akan diklasifikasikan membutuhkan variabel lain sebagai acuan untuk prediksi. jika tidak memasukkan variabel tersebut ke dalam algoritma svm, maka tidak akan ada acuan untuk memprediksikan data text tersebut. Dalam tugas akhir ini acuan yang digunakan untuk prediksi adalah variabel label. Oleh karena itu perlu untuk menyimpan variabel label dan membaginya menjadi *label train* dan *label test* [25]. Kemudian label yang sudah dibagi, akan digunakan sebagai variabel untuk pembuatan model klasifikasi menggunakan SVM.

4.2.4 Penggunaan metode *grid search*

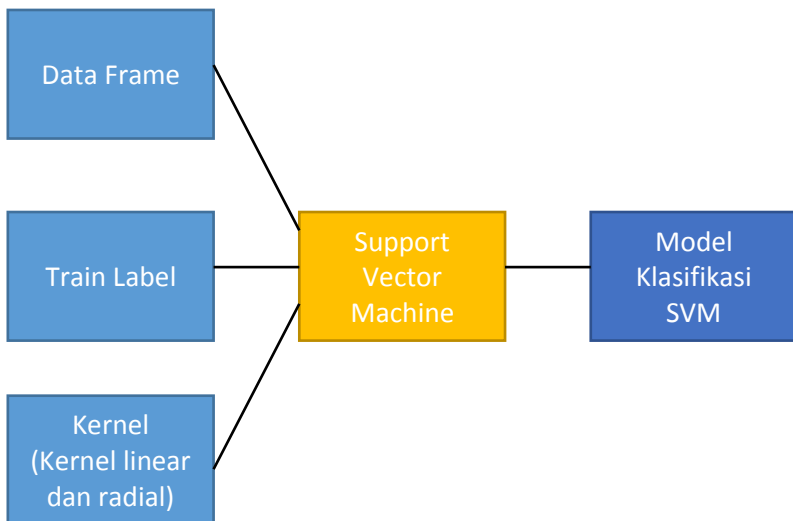
Dalam mencari model svm dengan akurasi tertinggi digunakan metode *grid search* untuk mencari nilai gamma dan cost terbaik untuk kernel radial dan mencari nilai cost terbaik untuk kernel linear. Dengan menggunakan metode ini, akan dilakukan pencarian parameter (baik nilai cost maupun gamma) satu per satu untuk menemukan parameter terbaik yang menghasilkan akurasi tertinggi dalam bentuk tabel [18].

Dalam tugas akhir ini penggunaan metode *grid search* dilakukan secara manual dengan menggunakan software R untuk mencari akurasi tertinggi dengan paramter yang telah ditentukan dan menggunakan Excel untuk membuat tabel untuk masing masing kernel *linear* dan *radial*.

4.2.5 Membuat Model Klasifikasi SVM

Model klasifikasi svm akan menggunakan variabel data frame train, label train dan kernel radial serta linear. Dalam penggunaan algoritma svm dalam r, dibutuhkan variabel [21] :

- x , yaitu data matriks, atau vektor yang merupakan data frame train pada tugas akhir ini.
- y , yaitu vector respon dengan satu label untuk setiap baris atau komponen x . dalam klasifikasi dapat berupa faktor yang merupakan train label pada tugas akhir ini.
- Kernel digunakan untuk *training* dan memprediksi. Untuk masing masing kernel yang digunakan dapat dirubah parameternya sesuai dengan jenis kernelnya. Pada kernel linear paramater yang dirubah adalah nilai *cost* sedangkan untuk kernel radial parameter yang dirubah adalah nilai *cost* dan *gamma*.



Gambar 4.3 Model Klasifikasi Support Vector Machine

BAB V IMPLEMENTASI

Bab ini menjelaskan proses pelaksanaan penelitian, implementasi klasifikasi menggunakan metode support vector machine.

5.1 Implementasi Data

Dalam pembuatan model klasifikasi, data yang digunakan berjumlah 886 Data dari 9680 data. Data yang digunakan hanya berjumlah 886 data, karena dari 9680 data yang didapat, data untuk kelas query dan non query tidak seimbang sehingga terjadi *imbalance dataset*. Dari 9680 data, data yang berjumlah data query sebanyak 443 dan sisanya adalah data non query.

Untuk mengatasi masalah *imbalance dataset* dapat dilakukan dengan metode *sampling*. Sampling dapat dicapai dengan 2 cara, yaitu dengan *under sampling the majority class*, *oversampling the minority class* atau *combining over and undersampling techniques*. Pada tugas akhir ini cara yang digunakan adalah *under sampling*. *Under sampling* merupakan metode untuk memecahkan masalah *imbalance dataset* dengan menghapus kelas yang dominan dalam data secara acak. Sehingga didapat jumlah data yang seimbang [26].

Oleh karena itu jumlah data query dan non query akan disamakan sehingga datanya seimbang. Sehingga data yang digunakan pada Tugas Akhir ini adalah 443 untuk masing masing kelas query dan non query dengan total data yang digunakan sebanyak 886 data.

Dari data dilakukan analisa statistik untuk mengetahui rata-rata, nilai max, min, dan standar deviasi dengan menghitung jumlah per kata yang ada pada teks. Dari hasil statistik dapat dilihat pada Tabel 5.1

Tabel 5.1 Hasil statistik

Rata – rata	7,019209
Maximun	139
Minimum	0
Standar Deviasi	9,821637

Dari tabel 5.1 kata terpendek adalah 0. Nilai 0 dikarenakan pada teks terdapat emoticon seperti “:)” yang dalam text mining, akan dihapus saat dilakukan pembersihan *corpus*. Untuk melihat panjang dari masing-masing teks yang ada pada data dapat dilihat pada Gambar 5.1. Data tersebut nantinya akan dibagi menjadi dua yaitu *train set* dan *test set* dengan perbandingan 70:30.

5.2 Proses Klasifikasi

Dalam melakukan klasifikasi menggunakan aplikasi R Studio, data akan dikelompokkan menjadi dua kelompok yaitu kelompok query dan non query. Metode yang digunakan dalam melakukan klasifikasi adalah support vector machine (SVM). Berikut merupakan implementasi dari model klasifikasi yang telah dijelaskan pada bab sebelumnya.

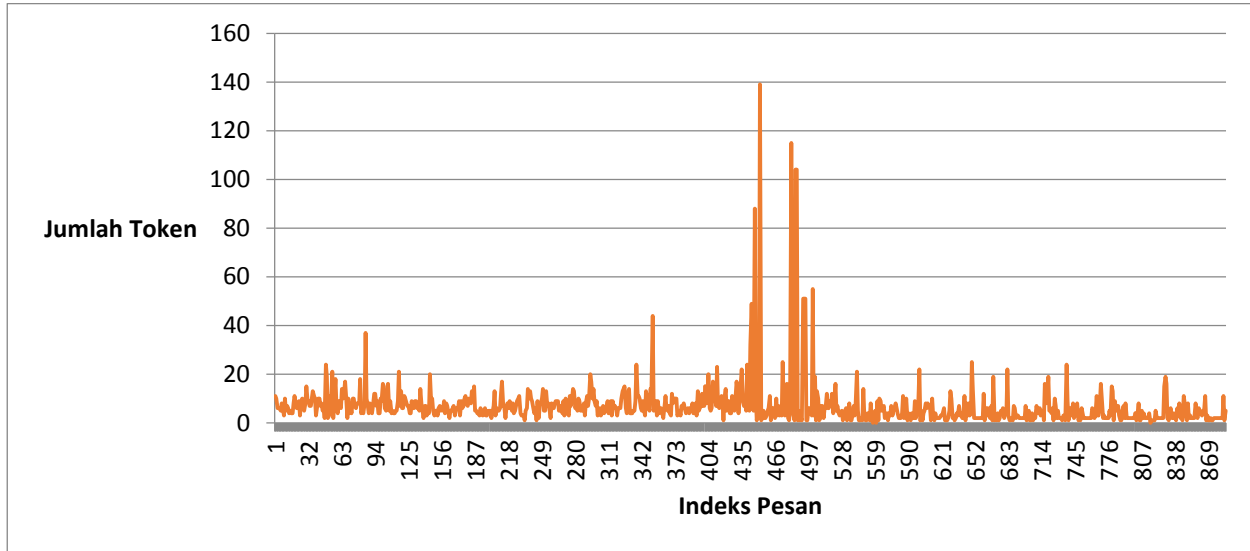
5.2.1 Menginputkan Data

Untuk melakukan input data pada R, data harus dalam bentuk .csv agar bisa dibaca oleh aplikasi. Setelah dimasukkan data yang sesuai, maka hasilnya seperti pada Kode 5.1.

```
datafb <-  
read.csv("~/Documents/TA/datafb.csv")
```

Kode 5.1 Input data pada R

Sebelum melakukan proses berikutnya, data yang sudah diinputkan dirubah kedalam bentuk *corpus*. seperti pada Kode 5.2.



Gambar 5.1 Grafik statistik data

```
#clean text
corpus_coba <- Corpus(VectorSource(datafb$text))
```

Kode 5.2 Merubah ke dalam bentuk *corpus*

5.2.2 Praproses Teks

Untuk melakukan praproses teks, dibutuhkan **library tm**.

Seperti yang sudah dijelaskan dalam bab perancangan , pada praproses teks, teks akan diubah ke dalam huruf kecil seperti pada Kode 5.3.

```
cleanset <-
tm_map(corpus_coba,content_transformer(tolowe
r))
```

Kode 5.3 Merubah ke dalam bentuk huruf kecil

Hasil dari data yang telah melalui proses ini dapat dilihat pada Tabel 5.2.

Tabel 5.2 Tabel hasil perubahan ke dalam huruf kecil

Teks Awal	Teks Setelah diubah ke dalam huruf kecil
Halo..salam kenal..aku mau pesan yg cetakan telur rebu	halo..salam kenal..aku mau pesan yg cetakan telur rebu

Setelah itu akan dilakukan penghapusan angka, dan tanda baca seperti pada Kode 5.4.

```
cleanset <- tm_map(cleanset, removeNumbers)
cleanset <- tm_map(cleanset, removePunctuation)
cleanset <- tm_map(cleanset, stripWhitespace)
```

Kode 5.4 Penghapusan angka dan tanda baca

Hasil dari data yang telah melalui proses ini dapat dilihat pada Tabel 5.3.

Tabel 5.3 Tabel hasil penghapusan angka dan tanda baca

Teks Awal	Teks Setelah dilakukan penghapusan angka dan tanda baca
Halo..salam kenal..aku mau pesan yg cetakan telur rebu	halosalam kenalaku mau pesan yg cetakan telur rebu

5.2.3 Menentukan Data Train dan Data Test

Setelah melalui tahap praproses teks, data yang sudah siap

```
#membagi data menjadi training dan testing, 70 : 30
size_data <- floor (0.7*nrow(datafb))
#set randomization seed
set.seed(141321)
indices_train <- sample(seq_len(nrow(datafb)),
  size = size_data)
data_train <- cleanset [indices_train ]
data_test <- cleanset [-indices_train ]
```

Kode 5.5 Membagi data menjadi data train dan data test

olah kemudian dibagi menjadi *data train* dan *data test* dengan perbandingan 70:30, seperti pada Kode 5.5.

5.2.4 Pembuatan DTM dan Data Frame

Data yang sudah dibagi kemudian perlu dirubah ke dalam bentuk *document term matrix* untuk masing masing *data train* dan *data test*.

Untuk membuat dtm seperti pada Kode 5.6.

```
#mengubah corpus menjadi document matrix di data
train dan test
train_dtm <- DocumentTermMatrix(data_train)
dtm_test <- DocumentTermMatrix(data_test, control
= list(dictionary = names(train_df)))
```

Kode 5.6 Pembuatan dtm untuk data train dan data test

Dalam pembuatan dtm, untuk menyamakan jumlah *terms* pada *data test* dan *data train*, maka dibuat *dictionary*. *Dictionary* yang dibuat berdasarkan kata yang paling sering muncul pada koleksi dokumen dalam *data train*. Kemudian untuk bisa masuk ke dalam klasifikasi menggunakan svm, dtm tersebut diubah ke dalam bentuk *data frame* seperti pada Kode 5.7.

```
train_df <-
  as.data.frame(data.matrix(train_dtm),
    stringsAsfactors = FALSE)
test_df <-
  as.data.frame(data.matrix(dtm_test),
    stringsAsfactors = FALSE)
```

Kode 5.7 Pembuatan data frame

5.2.5 Membuat Label Query dan Non Query

Pelabelan dilakukan untuk klasifikasi menggunakan svm pada aplikasi R Studio. Pelabelan dilakukan masing masing untuk *data train* dan *data test* seperti pada Kode 5.8.

```
train_label <- datafb$query [indices_train ]
test_label <- datafb$query [-indices_train ]
```

Kode 5.8 Pelabelan untuk data train dan data test

5.2.6 Klasifikasi menggunakan SVM

Dalam klasifikasi menggunakan svm dibutuhkan **library e1071** pada aplikasi R. Untuk mencari model svm yang tepat dilakukan klasifikasi terhadap data train dan label train serta

ditambahkan dengan parameter kernel linear atau radial seperti pada Kode 5.9.

```
Library(e1071)
model <- svm (train_df, train_label, kernel
= "linear")
model <- svm (train_df, train_label, kernel
= "radial")
```

Kode 5.9 Klasifikasi dengan svm

5.2.7 Penggunaan Metode *Grid Search*

Dalam mencari akurasi yang terbaik menggunakan kernel linear dan radial, perlu dilakukan percobaan dalam merubah parameter kernel untuk radial dan linear. Untuk kernel linear parameter yang diubah adalah nilai *cost* seperti pada Kode 5.10.

```
for(C in seq(-14, -13.3, by=0.1))
{
  modelsvm1 <- svm (train_df, train_label,
kernel = "linear", cost= 2^C)
  predicttrain1 <- predict(modelsvm1,
train_df)
  table(predicttrain1, train_label)
  predicttest1 <- predict(modelsvm1, test_df)
  table(predicttest1, test_label)
  print(100*sum(predicttest1==test_label)/len
gth(test_label))
}
```

Kode 5.10 Grid Search dengan kernel linear

Pada Kode 5.10, dilakukan pencarian nilai *cost* dengan rentang yang telah ditentukan. Kemudian outputan dari code di atas adalah hasil akurasi. Hasil akurasi dari nilai *cost* tersebut akan dibuat dalam bentuk tabel menggunakan excel. Kemudian untuk kernel *radial* parameter yang diubah adalah nilai *cost* dan *gamma* seperti pada Kode 5.11.

```
for(C in seq(-14, -13.3, by=0.1)){
  for(gamma in seq(0, -1, by=-0.5))
  {
    modelsvm1 <- svm (train_df,
                      train_label, kernel = "linear",
                      gamma = 2^gamma, cost= 2^C)
    predicttrain1 <- predict(modelsvm1,
                             train_df)
    table(predicttrain1, train_label)
    predicttest1 <- predict(modelsvm1,
                             test_df)
    table(predicttest1, test_label)
    print(100*sum(predicttest1==test_label)/length(test_label))
  }
}
```

Kode 5.11 Grid Search dengan Kernel Radial

Pada Kode 5.11, dilakukan pencarian nilai *cost* dan nilai *gamma* dengan rentang yang telah ditentukan. Kemudian outputan dari code di atas adalah hasil akurasi. Hasil akurasi dari nilai *cost* dan *gamma* tersebut akan dibuat dalam bentuk tabel menggunakan excel. Dari hasil tabel tersebut akan dilihat nilai parameter yang menghasilkan akurasi tertinggi.

5.2.8 Uji Model SVM

Kemudian untuk melakukan uji model menggunakan data test dan untuk mengukur hasil ketepatan dari model dalam memprediksikan kelas, diukur dengan menggunakan

confusion matrix untuk mendapatkan nilai akurasi, presisi, *recall* dan *F-Measure* seperti pada Kode 5.12.

```
predict <- predict(model, test_df)
100*sum(predict==test_label)/length(test_label)
confusionMatrix(test_label, predict)
```

Kode 5.12 Uji dan Mengukur model svm

5.3 Word Frequency Distribution

Word frequency distribution dilakukan untuk mengetahui distribusi dari kemunculan kata yang ada pada *corpus*. Dalam mencari distribusi dari frekuensi kata, dibuat pencarian dari 50 kata dengan frekuensi tertinggi pada *corpus*. Untuk melihat 50 distribusi frekuensi kata seperti pada Kode 5.13.

```
dtm <- DocumentTermMatrix(cleanset)
freq <- findFreqTerms(dtm, lowfreq = 30)
freq <- colSums(as.matrix(dtm))
length(freq)
ord <- order(freq, decreasing = TRUE)
freq[head(ord)]
freq[tail(ord)]
```

Kode 5.13 50 Distribusi frekuensi kata

Dari Kode 5.13 didapat hasil kata yang paling sering muncul dalam *corpus* lama (corpus sebelum dilakukan pembakuan kata) dapat dilihat pada Tabel 5.4.

Tabel 5.4 Top 50 Word Frequency Distribution *corpus* lama

No	Kata	Frekuensi Kemunculan
----	------	-------------------------

No	Kata	Frekuensi Kemunculan
1	ada	127
2	sis	119
3	brp	86
4	itu	79
5	subd	76
6	mold	62
7	cutter	61
8	mba	56
9	bento	49
10	nya	46
11	mbak	42
12	kalo	41
13	saya	39
14	nori	39
15	apa	38
16	mau	46
17	cetakan	34
18	rice	33
19	bear	32
20	klo	31
21	aja	31
22	total	28
23	egg	26
24	kurir	25
25	set	24
26	bisa	23

No	Kata	Frekuensi Kemunculan
27	face	23
28	<94><cf><dc><94><cf><a4>	22
29	buat	20
30	sandwich	20
31	sis	20
32	utk	20
33	yang	20
34	harga	19
35	hello	19
36	thx	19
37	ini	18
38	kitty	18
39	puncher	18
40	dan	16
41	sama	15
42	animals	14
43	bentuk	14
44	box	14
45	yah	14
46	aku	13
47	iya	13
48	kan	13
49	uda	13
50	angry	12

Dari Tabel 5.4 di atas diketahui terdapat kata yang mirip namun dengan penulisan yang berbeda, yaitu “kalo dan klo”,

“mba” dan “mbak”. Hasil dari grafik distribusi frekuensi kemunculan kata dengan frekuensi > 20 dapat dilihat pada Gambar 5.2. Untuk distribusi frekuensi seluruh kata pada corpus lama dapat dilihat pada Gambar 5.2. Dan untuk melihat seluruh distribusi kata pada *corpus* dapat dilihat pada Gambar 5.3.

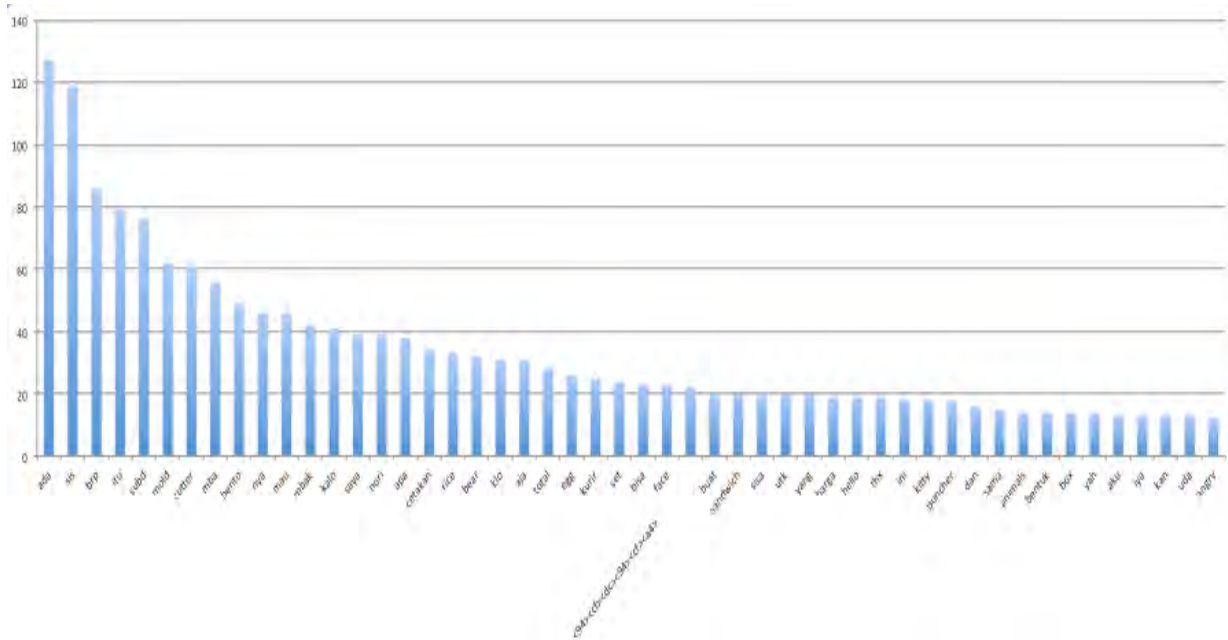
Dari hasil *word frequency distribution* maka dilakukan pembakuan kata, sehingga kata-kata yang memiliki makna yang sama namun penulisan yang berbeda akan dihapus dan diganti dengan kata-kata yang baku sehingga hanya terdapat satu kata. Pada Tabel 5.5 dapat dilihat kata yang dibakukan pada *corpus*. Kemudian dari hasil pembakuan kata tersebut, didapat *corpus* baru (*corpus* yang telah dibakukan). Berikut merupakan 50 kata yang paling sering muncul pada *corpus* yang baru dapat dilihat pada Tabel 5.6.

Dari *corpus* yang baru tersebut akan dibandingkan dengan *corpus* yang sebelumnya, apakah dengan dibakukannya kata akan menambahkan akurasi atau tidak. Selain itu dikarenakan pada 50 kata teratas dalam corpus baru dan corpus lama terdapat kata-kata yang termasuk ke dalam kata hubung seperti yang, maka akan dilakukan percobaan untuk menghapus *stopwords*. *Stopwords* yang digunakan merupakan *stopwords* bahasa indonesia [27]. Untuk menghapus *stopwords* dapat menggunakan Kode 5.14.

```
stopwordID <-
"http://raw.githubusercontent.com/nurandi/nurandi.net/master/data/stopwords-id.txt"
cstopwordID <- readLines(stopwordID);
cleanset <- tm_map(cleanset, removeWords,
c(cstopwordID))
```

Kode 5.14 *Remove Stopwords* Bahasa Indonesia

Dari hasil penghapusan stopwords pada corpus lama, 50 kata teratas pada corpus tersebut menjadi seperti pada Tabel 5.7.



Gambar 5.2 Grafik Distribusi Frekuensi Kata corpus lama

Tabel 5.5 Pembakuan kata pada corpus

Kata Baku	Kata dalam <i>Corpus</i>
Kalau	Klo, kalo, kalau, kl, klu, kalauu
Mbak	Mba, mbak, mbk, mb
Aku	Aq, aku, ak
Mold	Mold, mould, molds
Hello kitty	HK, hellokitty, hkitty
Atau	Ato, atau, or
Ada	Ada, ad
Ya	Y, ya, yaa
Lagi	Lagi, lg
Ini	Ini, ni
Berapa	Brp, berapa, brpa, brapa
Berapaan	brpan, brpaan, berapaan , brapaan
Beberapa	Bbrp, beberapa
Sis	Sis, sist, ssis, siss
Untuk	Utk, untuk
Juga	Juga, jg, jg
Punya	Punya, pny, pnyk
Yang	Yg, yang
Sama	Sama, ama, sma, sm
Saya	Saya, sy, sya, ssy
Minta	Minta, mnta
Tidak	Enggk, gak , ga, nga , nda, tidak
Maaf	Maaf, maap, sorry, sory
Masih	Masih, msh
Harga	Harga, hrg, hrga, hargag
Bisa	Bisa, bs,
Harganya	hrgnya, harganya,
Lagi	Lagi, lg, laagi
Terima kasih	Terima kasih, thanks, thx, tengkyu, trims, tks, trmkash
Tolong	Tlg, tlong

Kata Baku	Kata dalam <i>Corpus</i>
Gambar	Gbr, gbrnya
Ada	Ada, ad
Terus	Trus, terus
Kayak	Kayak, kyk,
Pesan	Pesen, psn
Kemarin	Kmren, kmrn,
Nya	Nya, ny,
Gitu	Gitu, gitu, gt
Hari	Hari, hr
Jadi	Jadi, jd
Bisa	Bisa, bs, bsa
Pakai	Pake, pk, pke
Masuk	Masuk, msuk
Sudah	Uda, sudah, udah, sdh, dah, sdah, uuda
Tapi	Tapi, tp
Dulu	Dulu, dl
Minta	Minta, mnt
Terus	Terus, trs
Seperti	Seperti, sp
Dapat	Dpt
Ok	Ok, oke, k, okay
Transfer	transf, transfer, tt
Bayar	Bayar, byr, byar
Tadi	Tadi, td
Tanya	Tanya, tnya, tny

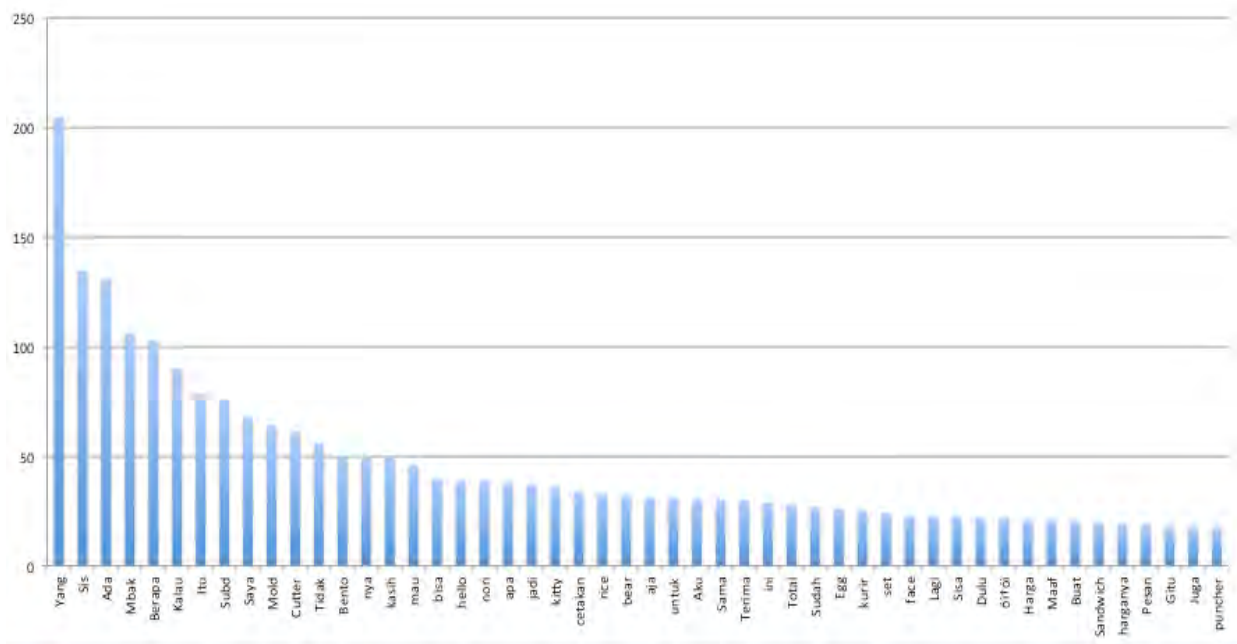
Tabel 5.6 Top 50 Word Frequency Distribution *corpus* lama

No	Kata	Frekuensi Kemunculan
1	Yang	205
2	Sis	135
3	Ada	131

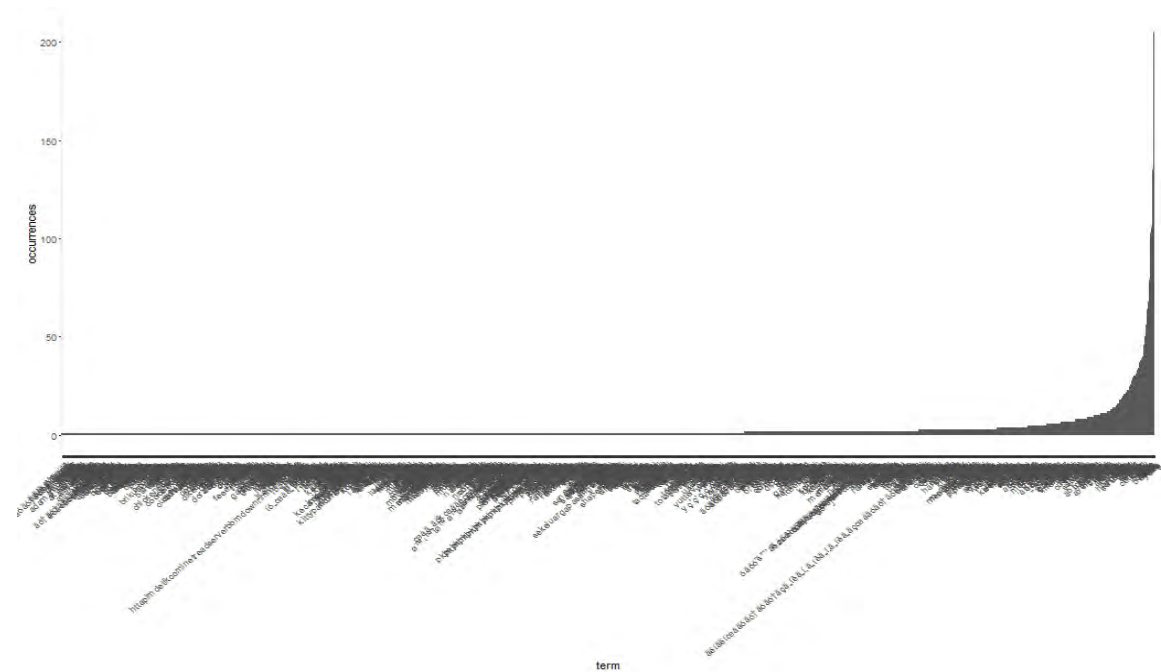
No	Kata	Frekuensi Kemunculan
4	Mbak	106
5	Berapa	103
6	Kalau	90
7	Itu	79
8	Subd	76
9	Saya	68
10	Mold	64
11	Cutter	61
12	Tidak	56
13	Bento	50
14	nya	49
15	mau	46
16	bisa	40
17	hello	39
18	kasih	49
19	nori	39
20	apa	38
21	jadi	37
22	kitty	36
23	cetakan	34
24	rice	33
25	bear	32
26	aja	31
27	untuk	31
28	Aku	30
29	Sama	30

No	Kata	Frekuensi Kemunculan
30	Terima	30
31	ini	29
32	Total	28
33	Sudah	27
34	Egg	26
35	kurir	25
36	set	24
37	face	23
38	Lagi	23
39	Sisa	23
40	Dulu	22
41	óì†òì	22
42	Harga	21
43	Maaf	21
44	Buat	20
45	Sandwich	20
46	harganya	19
47	Pesan	19
48	Gitu	18
49	Juga	18
50	puncher	18

Untuk melihat grafik frekuensi kemunculan kata dari *corpus* yang baru dapat dilihat pada **Error! Reference source not found.2.**



Gambar 5.4 Grafik Distribusi Frekuensi Kata *corpus* baru



Gambar 5.5 Gambar distribusi seluruh kata *corpus* baru

Tabel 5.7 50 Kata Teratas pada corpus lama

No	Kata	Frekuensi Kemunculan
1	sis	119
2	brp	86
3	subd	76
4	mold	62
5	cutter	61
6	mba	56
7	bento	49
8	nya	46
9	mbak	42
10	kalo	41
11	nori	39
12	cetakan	34
13	rice	33
14	bear	32
15	aja	31
16	klo	31
17	total	28
18	egg	26
19	kurir	25
20	set	24
21	face	23
22	îœ<îœ	22
23	sandwich	20
24	sis	20
25	utk	20

No	Kata	Frekuensi Kemunculan
26	harga	19
27	hello	19
28	thx	19
29	kitty	18
30	puncher	18
31	animals	14
32	bentuk	14
33	box	14
34	yah	14
35	iya	13
36	uda	13
37	angry	12
38	isi	12
39	pcs	12
40	ready	12
41	anak	11
42	ayahbunda	11
43	bayar	11
44	gitu	11
45	msh	11
46	pan	11
47	punya	11
48	rabbit	11
49	ricemold	11
50	sist	11

Sedangkan unuk 50 kata teratas pada corpus baru setelah dilakukan penghapusan stopwords seperti pada Tabel 5.7

Tabel 5.8 50 Kata Teratas pada *corpus* baru

No	Kata	Frekuensi Kemunculan
1	sis	134
2	mbak	115
3	subd	76
4	mold	64
5	cutter	62
6	nya	50
7	bento	49
8	kitty	42
9	hello	40
10	nori	39
11	cetakan	34
12	aja	33
13	bear	33
14	rice	32
15	terimakasih	29
16	total	29
17	untuk	27
18	egg	26
19	kurir	25
20	set	25
21	”ü”i	22
22	face	22
23	harga	22
24	sis	22
25	jadi	21
26	sandwich	21
27	harganya	19

No	Kata	Frekuensi Kemunculan
28	gitu	18
29	hari	18
30	pesan	18
31	puncher	17
32	maaf	16
33	iya	15
34	animals	14
35	box	14
36	mint	14
37	uang	14
38	angry	13
39	bayar	13
40	bentuk	13
41	tolong	13
42	yah	13
43	%\u008dbiåå	12
44	pcs	12
45	punya	12
46	ready	12
47	anak	11
48	ayahbunda	11
49	isi	11
50	pan	11

Dari hasil penghapusan stopwords akan dilakukan pencarian nilai akurasi apakah mengurangi atau menambah akurasi untuk masing-masing corpus lama dan baru.

BAB VI

UJI COBA DAN ANALISIS HASIL

Bab ini berisikan hasil dan pembahasan setelah melakukan implementasi. Hasil yang akan dijelaskan adalah hasil uji coba model, pembahasan tentang hal yang menyebabkan hasil yang ada terjadi.

6.1 Membuat Model Uji Coba

Pada tahapan ini akan dibuat model uji coba pada *train* set dengan beberapa jenis model untuk bisa melakukan perbandingan dalam menentukan model terbaik yang akan digunakan untuk klasifikasi. Skenario uji coba dapat dilihat pada Tabel 6.1.

Tabel 6.1 Skenario Uji Coba

Uji Coba	Skenario
I	Model dirancang menggunakan parameter <i>kernel linear</i> dengan nilai <i>cost</i> dari 2^{-19} sampai dengan 2^{10} dengan rentang 1.
II	Model dirancang menggunakan parameter <i>kernel linear</i> dengan nilai <i>cost</i> dari 2^{-14} sampai dengan 2^{-11} dengan rentang 0.1
III	Model dirancang menggunakan parameter <i>kernel radial</i> dengan nilai <i>gamma</i> dari 2^{-19} sampai dengan 2^5 dengan rentang 1 dan dengan nilai <i>cost</i> dari $2^{-1.55}$ sampai dengan $2^{-1.05}$ dengan rentang -0.05, dan dari 2^{-1} sampai dengan 2^1 dengan rentang 0.5, dan dari $2^{1.05}$ sampai dengan $2^{1.55}$ dengan rentang 0.05.

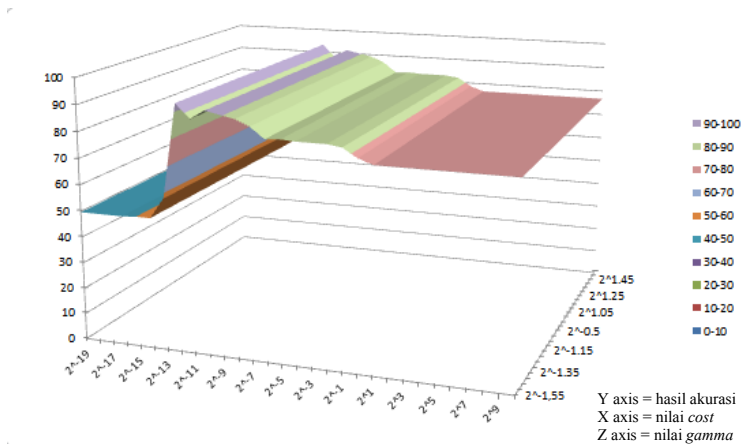
Uji Coba	Skenario
IV	Model dirancang menggunakan parameter <i>kernel radial</i> dengan nilai <i>gamma</i> dari 2^{-17} sampai dengan 2^{-14} dengan rentang 0.1 dan nilai <i>cost</i> dari $2^{1.00}$ sampai dengan $2^{1.55}$ dengan rentang 0.01.
V	Model dirancang dengan melakukan pembakuan kata dari kata kata yang memiliki makna yang sama namun dengan penulisan kata yang berbeda. Banyaknya kata yang dibakukan dapat dilihat pada Tabel 11. Pada uji coba ini dibuatlah <i>corpus</i> baru yang didalamnya sudah terdapat kata yang di bakukan.
VI	Model dirancang dengan melakukan penghapusan stopwords berbahasa indonesia untuk corpus lama dan corpus baru.

6.1.1 Uji Coba I

Pada Uji Coba I, model dirancang dengan menggunakan parameter kernel linear dengan nilai *cost* 2^{-19} sampai dengan 2^{10} dengan rentang 1. Dengan menggunakan metode *grid search* dilakukan pencarian nilai *cost* akurasi tertinggi. Pada aplikasi R Studio, saat dilakukan pencarian nilai *cost* menggunakan kernel linear diketahui bahwa terdapat nilai *gamma*, walaupun dalam pembuatan model hanya memasukkan nilai *cost*. Akan tetapi setelah dilakukan percobaan dengan memasukkan nilai *gamma*, ternyata tidak berpengaruh terhadap model yang dibuat. Diagram diagram dari perancangan model dengan nilai *cost* yang telah ditentukan dapat dilihat pada Gambar 6.1.

Pada Gambar 6.1 Y-axis menunjukkan hasil akurasi, X-axis menunjukkan nilai *cost* dan Z-axis menunjukkan nilai *gamma*. Dari Gambar 6.1 diketahui bahwa nilai *cost* untuk nilai yang negati dari rentang 2^{-19} sampai dengan 2^{-13}

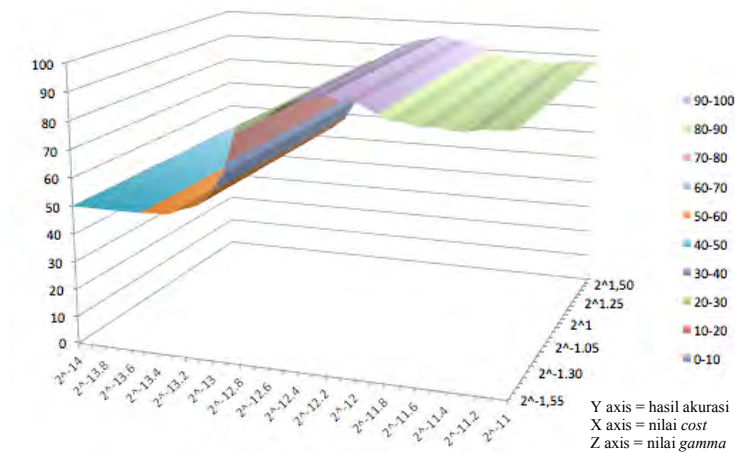
memiliki nilai akurasi yang kurang bagus, namun dari rentang 2^{10} memiliki nilai akurasi yang lebih bagus.



Gambar 6.1 Grafik KeUji Coba II

Pada Uji Coba II, model dirancang dengan menggunakan parameter kernel linear dengan nilai *cost* mulai dari 2^{-14} sampai dengan 2^{-11} dengan rentang 0.1. Dengan menggunakan metode grid search dilakukan pencarian nilai *cost* yang menghasilkan nilai akurasi tertinggi. Diagram hasil perancangan model dengan nilai *cost* yang telah ditentukan dapat dilihat pada Gambar 6.2. Pada Gambar 6.2 .

Dari Gambar 6.2 diketahui bahwa nilai *cost* untuk nilai yang negatif dari rentang 2^{-14} sampai dengan $2^{-12.9}$ memiliki nilai akurasi yang kurang bagus, namun dari rentang $2^{-12.8}$ sampai dengan 2^{-11} memiliki nilai akurasi yang lebih bagus.

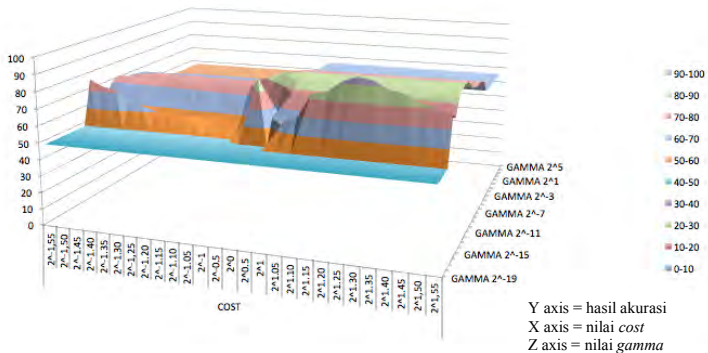
Gambar 6.2 Grafik *Kernel Linear 1*

6.1.2 Uji Coba III

Pada uji coba III, model dirancang dengan menggunakan parameter kernel radial dengan nilai *gamma* 2^{-19} sampai dengan 2^5 dengan rentang 1 dan *cost* dengan nilai $2^{-1.55}$ sampai dengan $2^{-1.05}$ dengan rentang -0.05, dan 2^{-1} sampai dengan 2^1 dengan rentang 0.5, dan $2^{1.05}$ sampai dengan $2^{1.55}$ dengan rentang 0.05. Dengan menggunakan metode *grid search* dilakukan pencarian pasangan *cost* dan *gamma* yang menghasilkan nilai akurasi tertinggi. Diagram hasil perancangan model dengan pasangan *cost* dan *gamma* yang telah ditentukan dapat dilihat pada Gambar 6.3.

Dari Gambar 6.3 diketahui bahwa nilai *cost* untuk nilai yang negatif memiliki akurasi kurang bagus dibandingkan dengan nilai *cost* yang positif. Selain itu nilai *gamma* untuk nilai yang positif memiliki akurasi yang kurang bagus dibandingkan dengan nilai *gamma* yang negatif. Dari uji coba model ini dihasilkan nilai akurasi yang bagus dengan

nilai gamma sebesar 2^{-15} dengan nilai cost positif dari $2^{1.10}$ sampai dengan $2^{1.55}$.

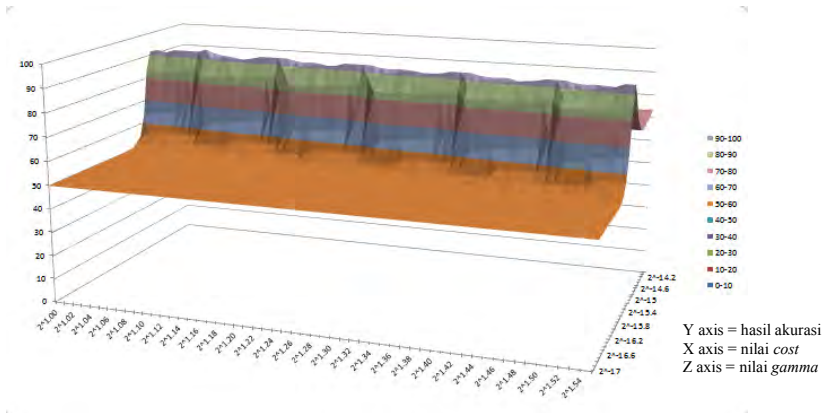


Gambar 6.3 Grafik *Kernel Radial 1*

6.1.3 Uji Coba IV

Pada uji coba IV, model dirancang menggunakan parameter kernel radial dengan nilai gamma 2^{-17} sampai dengan 2^{-14} dengan rentang 0.1 dan nilai cost $2^{1.00}$ sampai dengan $2^{1.55}$ dengan rentang 0.01. Dengan menggunakan metode *grid search* dilakukan pencarian pasangan *cost* dan *gamma* yang menghasilkan nilai akurasi tertinggi. Diagram dari hasil perancangan model dengan pasangan *cost* dan *gamma* yang telah ditentukan dapat dilihat pada Gambar 6.4.

Dari Gambar 6.4 diketahui bahwa nilai *gamma* dengan rentang $2^{-15.1}$ sampai dengan $2^{-14.8}$ memiliki nilai akurasi yang lebih bagus. Nilai *cost* yang memiliki akurasi yang bagus didapat pada nilai $2^{1.03}$ sampai dengan $2^{1.06}$, $2^{1.14}$ sampai dengan $2^{1.17}$, $2^{1.24}$ sampai dengan $2^{1.27}$ dan $2^{1.36}$ sampai dengan $2^{1.38}$

Gambar 6.4 Grafik *Kernel Radial*

6.1.4 Uji Coba V

Pada uji coba V dilakukan pembakuan kata dari kata kata yang memiliki makna yang sama namun dengan penulisan kata yang berbeda. Banyaknya kata yang dibakukan dapat dilihat pada Tabel 11. Pada uji coba ini dibuatlah *corpus* baru yang didalamnya sudah terdapat kata yang di bakukan. Kemudian hasil dari uji coba ini, akan dibandingkan dengan corpus sebelumnya apakah menambah akurasi atau mengurangi akurasi. Dari hasil uji coba ini diketahui bahwa pembakuan kata tidak menambahkan akurasi namun mengurangi hasil akurasi.

6.1.5 Uji Coba VI

Pada uji coba VI dilakukan penghapusan *stopwords* bahasa indonesia untuk masing-masing corpus lama dan *corpus* baru. Dari hasil uji coba ini juga dilakukan analisis *corpus* dengan mencari 50 kata teratas dari masing-masing *corpus*. Hasil uji coba menunjukkan bahwa dengan dilakukan penghapusan *stopwords* bahasa indonesia cenderung menurunkan nilai akurasi, dan

parameter terbaik dengan akurasi tertinggi adalah *kernel linear* dengan nilai $cost\ 2^{-12}$.

6.2 Hasil Uji Coba Model

Dari hasil uji coba model diketahui bahwa :

- pada uji coba I nilai akurasi sebesar 93.23% didapat dengan nilai $cost\ 2^{-12}$.
- Pada uji coba II nilai akurasi sebesar 94.74% didapat dengan nilai $cost\ 2^{-12.2}$.
- Pada uji coba III nilai akurasi sebesar 92.48% didapat dengan nilai $cost\ 2^{-1.25}$ dan nilai $gamma\ 2^{-15}$.
- Pada uji coba IV nilai akurasi sebesar 92.86% didapat dengan pasangan nilai $gamma\ 2^{-14.8}$ dan $cost\ 2^{-1.03}$ sampai dengan $2^{-1.06}$, nilai $gamma\ 2^{-14.9}$ dan $cost\ 2^{-1.14}$ sampai dengan $2^{-1.17}$, nilai $gamma\ 2^{-15}$ dan $cost\ 2^{-1.24}$ sampai dengan $2^{-1.27}$, nilai $gamma\ 2^{-15.1}$ dan $cost\ 2^{-1.36}$ sampai dengan $2^{-1.38}$.
- Pada uji coba V, diketahui bahwa nilai akurasi berdasarkan parameter kernel radial dan linear dapat dilihat pada Tabel 6.2.

Tabel 6.2 Nilai akurasi berdasarkan parameter

Parameter	Hasil Akurasi
<i>Kernel linear</i> dengan nilai $cost\ 2^{-12}$	90.60 %
<i>Kernel linear</i> dengan nilai $cost\ 2^{-12.2}$	92.48%
<i>Kernel radial</i> dengan nilai $cost\ 2^{-1.25}$ dan $gamma\ 2^{-15}$	90.60%
<i>Kernel radial</i> dengan nilai $cost\ 2^{-1.03}$ dan $gamma\ 2^{-14.8}$	87.21%

Dari hasil uji coba V, diketahui nilai akurasi tertinggi adalah dengan menggunakan kernel linear dengan nilai $cost\ 2^{-12.2}$.

- Pada uji coba VI, diketahui nilai akurasi berdasarkan parameter kernel radial dan linear untuk corpus lama dapat dilihat pada Tabel 6.3

Tabel 6.3 Nilai Akurasi *corpus* lama

Parameter	Hasil Akurasi
<i>Kernel linear</i> dengan nilai $cost\ 2^{-12}$	91.72 %
<i>Kernel linear</i> dengan nilai $cost\ 2^{-12.2}$	86.09%
<i>Kernel radial</i> dengan nilai $cost\ 2^{1.25}$ dan $gamma\ 2^{-15}$	83.83%
<i>Kernel radial</i> dengan nilai $cost\ 2^{1.03}$ dan $gamma\ 2^{-14.8}$	83.83%

Sedangkan nilai akurasi untuk corpus baru dapat dilihat pada Tabel 6.4.

Tabel 6.4 Nilai Akurasi *corpus* baru

Parameter	Hasil Akurasi
<i>Kernel linear</i> dengan nilai $cost\ 2^{-12}$	90.97%
<i>Kernel linear</i> dengan nilai $cost\ 2^{-12.2}$	88.72%
<i>Kernel radial</i> dengan nilai $cost\ 2^{1.25}$ dan $gamma\ 2^{-15}$	81.57%
<i>Kernel radial</i> dengan nilai $cost\ 2^{1.03}$ dan $gamma\ 2^{-14.8}$	81.57%

Dari hasil uji coba VI, diketahui bahwa untuk corpus yang lama nilai akurasi tertinggi adalah 91.72% menggunakan kernel linear dengan nilai cost adalah 2^{-12} . Sedangkan untuk corpus yang baru nilai akurasi tertinggi adalah 90.97% menggunakan kernel linear dengan nilai cost adalah 2^{-12} .

Kemudian dari hasil uji coba untuk masing masing model akan dibandingkan untuk menentukan model terbaik yang akan digunakan untuk klasifikasi. Pada Tabel 6.5 dapat dilihat hasil akurasi untuk masing masing model uji coba dengan skenario yang berbeda.

Tabel 6.5 Tabel akurasi hasil uji coba model

Model Uji Coba	Hasil Akurasi
Uji Coba 1	93.23%
Uji Coba II	94.74%
Uji Coba III	92.48%
Uji Coba IV	92.86%
Uji Coba V	92.48%
Corpus Lama	91.72%
Corpus Baru	90.97%

Dari Tabel 6.5 diketahui bahwa nilai model uji coba II memiliki nilai akurasi tertinggi sehingga menjadi model yang terbaik yang akan digunakan untuk klasifikasi.

6.3 Uji Validasi

Uji Validasi dilakukan untuk mengevaluasi model klasifikasi. Evaluasi dari model klasifikasi diukur berdasarkan perhitungan akurasi, presisi, *recall* dan *F-Measure*. Akurasi merupakan kinerja model dalam mengklasifikasikan teks,

presisi merupakan keakuratan model, recall merupakan sensitivitas model, dan F-Measure merupakan kemampuan model dalam menggali informasi teks.

Pada bagian ini akan ditampilkan hasil uji validasi model dengan metode SVM, penghapusan beberapa kata, dan penggunaan *grid search* untuk mencari nilai parameter gamma dan cost terbaik untuk kernel linear maupun radial. Hasil uji validasi model dapat dilihat pada Tabel 6.6.

Tabel 6.6 Tabel Hasil Uji Validasi Model

Skenario	Akurasi	Presisi	Recall	F-Measure
Uji Coba I	93.23%	89.39%	96.72%	93.52%
Uji Coba II	94.74%	93.18%	96.09%	96.18%
Uji Coba III	92.86%	94.70%	91.24%	92.93%
Uji Coba IV	92.48%	95.45%	90%	92.64%
Uji Coba V	92.48%	88.64%	95.90%	92.12%
Corpus Lama	91.72%	91.67%	91.67%	91.67%
Corpus Baru	90.97%	92.42%	89.71%	91.04%

Berdasarkan Tabel 6.6 diketahui bahwa hasil akurasi tertinggi adalah Uji Coba II. Dari enam percobaan, pada percobaan I – IV tidak dilakukan penghapusan kata pada *corpus*. Sedangkan pada percobaan V dilakukan pembakuan kata pada *corpus* dan pada percobaan VI dilakukan penghapusan *stopwords*. Jadi dapat dikatakan bahwa dengan dilakukannya pembakuan kata akan mengurangi akurasi. Sedangkan kebalikannya akurasi yang didapatkan lebih tinggi. Selain itu dengan penghapusan *stopwords* pada masing-masing corpus akan mengurangi akurasi.

6.4 Analisis Hasil Uji Coba Model

6.4.1 Analisis Uji Validasi

Hasil uji validasi terhadap model klasifikasi dengan menggunakan metode Support Vector Machine (SVM) mencakup perhitungan akurasi, presisi, *recall*, dan *F-Measure*. Dari hasil uji validasi model diketahui bahwa model klasifikasi terbaik adalah model uji coba II yang memiliki akurasi di atas 90%. Dari hasil uji diketahui bahwa nilai akurasi model adalah sebesar 94.74%, presisi adalah 93.18%, *recall* adalah 96.09% dan *F-Measure* adalah 96.18%.

Berdasarkan hasil uji validasi maka hasil evaluasi dari uji coba model diketahui bahwa nilai akurasi adalah 94.74%. Dengan nilai akurasi tersebut maka dapat dikatakan bahwa dengan kinerja model klasifikasi sudah bagus. Nilai presisi dari hasil evaluasi model adalah 93.18%. Dengan nilai presisi tersebut dapat dikatakan bahwa model bagus dalam mengkategorikan teks ke dalam kelas yang seharusnya, sehingga dengan kata lain keakuratan model adalah baik. Nilai *recall* dari hasil evaluasi model adalah 96.09%. Dengan nilai *recall* tersebut dapat dikatakan bahwa model sudah bagus dalam memprediksikan teks dengan benar, sehingga dengan kata lain model memiliki sensitivitas yang baik. Nilai *F-Measure* dari hasil evaluasi model adalah 96.18%. dengan nilai *F-Measure* tersebut dapat dikatakan bahwa model klasifikasi dapat menggali informasi teks dengan baik.

Namun dengan akurasi yang baik, belum menunjukkan performa model yang baik. Hal ini dikarenakan dalam perhitungan akurasi, jika distribusi kelas tidak merata maka hasil perhitungan akurasi bisa saja didapat dari memprediksikan kelas yang dominan sehingga menghasilkan nilai akurasi yang tinggi namun menghasilkan presisi dan *recall* yang rendah untuk kelas yang lain.

Oleh karena itu untuk mengukur performa dari setiap kelas yang ada pada *dataset* maka akan dilakukan perhitungan

presisi, *recall*, dan *F-Measure* dari masing-masing kelas. Hasil perhitungan presisi, *recall*, dan *F-Measure* dari masing-masing kelas dapat dilihat pada Tabel 6.7.

Tabel 6.7 Tabel perhitungan presisi, recall, F-Measure masing-masing kelas

Kelas	Presisi	Recall	F-Measure
Query	93.18%	96.09%	96.18%
Non-Query	96.27%	93.48%	94.85%

Berdasarkan Tabel 6.7 diketahui bahwa berdasarkan tiga parameter (presisi, *recall* dan *F-Measure*) menghasilkan nilai di atas 80% untuk kelas query dan non query. Hal ini menunjukkan bahwa performa model cukup baik dalam mengklasifikasikan kelas query dan non query. Dari hasil akurasi model maupun presisi, *recall*, dan *F-Measure* dari masing – masing kelas, dapat dikatakan bahwa hasil yang didapatkan dan performa *Support Vector Machine* (SVM) dalam mengklasifikasin teks permintaan informasi adalah baik.

6.4.2 Analisis Perbandingan Uji Coba

Berdasarkan hasil uji coba VI dengan melakukan pembakuan kata dibandingkan dengan uji coba II yang menggunakan data awal tanpa dilakukan pembakuan kata. Hasil perbandingan akurasi dapat dilihat pada Tabel 6.8

Tabel 6.8 Perbandingan akurasi model uji coba II dan VI

Model Uji Coba	Akurasi	Presisi	Recall	F-Measure
II	94.74%	93.18%	96.09%	96.18%
VI	92.48%	88.64%	95.90%	92.12%

Berdasarkan hasil akurasi di atas, maka diketahui nilai akurasi tertinggi didapat dari model uji coba II dengan nilai

94.74%. Untuk mengetahui masing-masing kelas yang diprediksikan benar dapat dilihat pada Tabel 6.9

Tabel 6.9 *Confusion Matrix* model II

Predict test	Test label	
	Non query	query
Non query	129	9
Query	5	123

Pada uji coba II diketahui teks pada kelas non query yang diprediksikan benar sebanyak 129 dan diprediksikan salah sebanyak 5. Sedangkan teks pada kelas query yang diprediksikan benar sebanyak 123 dan diprediksikan salah sebanyak 9. Dalam uji coba ini, kata yang diprediksikan salah pada kelas query dapat dilihat pada Tabel 6.10

Tabel 6.10 Teks yang diprediksi salah pada kelas query

No	Teks	Kelas pada model	Kelas pada data asli
19	maaf, mbak.. saya sudah nemu di Bdg, tidak jadi pesan jadinya maaf, mbak.. saya sudah nemu di Bdg, tidak jadi pesan jadinya	Non-query	Query
27	Mbak kalau pmbelanjaan brktnya disc berapa ya?	Non-query	Query
51	Semua ukurannya xxl?	Non-query	Query
84	Nanti <e5><bb>k<ec><dc> <95><a3><c7><89><cf><e4><cc><d9><cc>	Non-query	Query

No	Teks	Kelas pada model	Kelas pada data asli
	><d9><89><aa><8f><95><a3><c7><e5><a8> > barang Y<e5><bb><eb><a8> g <e5><bb>k<ec><dc> ambil sis...		
204	yang nori pounch apa <89><db>_<89><db> <dc><94>_<a5>G<ed> >_<ed><a8><ed><c9> <ed>_<ea><c1><cc><b4> <bc><ed>_<ed><a8><ed><c9><ed>_<ea><c1><94>_<a5> kurang sis?	Non-query	Query
226	mbak bear n rabbitku + rm segitiga dikirim senin ya	Non-query	Query
227	Punchernya kali ini bisa lbh murah dikit soalnya naik kapal	Non-query	Query
231	Dv-ds-fw 1 ya	Non-query	Query
246	Cp-tr-bk sis	Non-query	Query

Berdasarkan Tabel 6.10, diketahui bahwa kesalahan dalam klasifikasi bukanlah kesalahan model dalam mengklasifikasikan namun, karena terdapat teks yang seharusnya termasuk ke dalam kelas non-query namun diberi label query, seperti pada no 19, 27, 51, 84, 227, 231, dan 246. Sedangkan kata yang diprediksikan salah pada kelas non query dapat dilihat pada Tabel 6.11.

Tabel 6.11 Teks yang diprediksi salah pada kelas non-query

No	Teks	Kelas pada model	Kelas pada data asli
468	Sama yg flower	Query	Non-query
485	<p>Ayo ketemu komunitas AYAHBUNDA bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREAST'</p> <p>Minggu - 8 Juli 2012</p> <p>Pk. 09.30 - 13.00</p> <p>WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya</p> <p>Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila</p> <p>Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda)</p> <p>Dress code : touch of gold</p> <p>50 peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan bingkisan menarik.</p>	Query	Non-query

	<p>Ticket Box : ANGELINA - 0856 312 7797 Pin BB : 20EB14C0 Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -</p>		
486	<p>Ayo ketemu komunitas AYAHBUNDA bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREASI' Minggu - 8 Juli 2012 Pk. 09.30 - 13.00 WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda) Dress code : touch of gold 50 peserta pertama yang datang sebelum pk. 09.30 akan</p>	Query	Non-query

	mendapatkan bingkisan menarik. Ticket Box : ANGELINA - 0856 312 7797 Pin BB : 20EB14C0 Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -		
524	Mba puncher face yg murmer <80><c9><ec><d4>< 82> <81><e4><ae><8d><a e><8d><e5><bb><80 ><fc>? :D	Query	Non-query
569	Kirimnya brp paket?	Query	Non-query

Pada Tabel 6.11 diketahui bahwa terdapat teks yang termasuk ke dalam kelas query namun diberi label non query seperti pada nomor 468 dan 524. Sedangkan untuk nomor 485, 486 dan 569 merupakan teks yang termasuk ke dalam kelas non query namun diprediksi kelas query.

Untuk uji coba V diketahui bahwa nilai akurasi adalah 92.48%. akurasi yang dihasilkan lebih rendah daripada uji coba II. Untuk mengetahui masing-masing kelas yang diprediksikan benar dapat dilihat pada Tabel 6.12.




































Tabel 6.12 Confusion Matrik model Uji Coba V

Predict test	Test label	
	Non query	query
Non query	129	15
Query	5	117

Pada uji coba V diketahui teks pada kelas non query yang diprediksikan benar sebanyak 129 dan diprediksikan salah sebanyak 5. Sedangkan teks pada kelas query yang diprediksikan benar sebanyak 117 dan diprediksikan salah sebanyak 15. Dalam uji coba ini, kata yang diprediksikan salah pada kelas query dapat dilihat pada Tabel 6.13

Tabel 6.13 Teks yang diprediksikan salah pada kelas query

No	Teks	Kelas pada model	Kelas pada data asli
19	maaf, mbak.. saya sudah nemu di Bdg, tidak jadi pesan jadinya maaf, mbak.. saya sudah nemu di Bdg, tidak jadi pesan jadinya	Non-query	Query
27	Mbak kalau pmbelanjaan brktnya disc berapa ya?	Non-query	Query
51	Semua ukurannya xxl?	Non-query	Query
56	maaf sis, saya kurang suka. kalau puncher yang kosong saya inden bisa?	Non-query	Query
67	Dv-ds-fw 1 ya	Non-query	Query

No	Teks	Kelas pada model	Kelas pada data asli
84	Nanti   k              /△       barang Y   g   k     ambil sis...	Non- query	Query
135	berapa hari itu lesnya mbak	Non- query	Query
199	Pancake ring chicken,fish,rabbit dan love	Non- query	Query
226	mbak bear n rabbitku + rm segitiga dikirim senin ya	Non- query	Query
227	Punchernya kali ini bisa lbh murah dikit soalnya naik kapal	Non- query	Query
231	Dv-ds-fw 1 ya	Non- query	Query
246	Cp-tr-bk sis	Non- query	Query
272	Cetakan norinya bkn 30rb mbak:D	Non- query	Query
343	Eh di pp saya ricemold kepala bunny sama bear	Non- query	Query
358	Nori juga   麴 a     _ di forbento ?	Non- query	Query

Berdasarkan pada Tabel 6.13, diketahui bahwa kesalahan dalam klasifikasi bukanlah kesalahan model dalam mengklasifikasikan namun, karena terdapat teks yang

seharusnya termasuk ke dalam kelas non-query namun diberi label query, seperti pada no 19, 27, 51, 67, 84, 231, 246. Tetapi untuk nomor 56, 135, 199, 226, 227, 272, 343, dan 358 merupakan kelas query yang diprediksi non query. Untuk mengetahui kata yang menyebabkan model salah memprediksi maka pada Tabel 6.14 dilakukan perbandingan perubahan kata dari model lama yang diprediksi benar dan diprediksi salah pada model baru. Dari Tabel 6.14 dapat dilihat bahwa hasil pembakuan kata ternyata membuat model salah memprediksikan kata sehingga banyaknya teks yang diprediksi salah pada model baru lebih banyak dari pada model lama. Untuk teks yang diprediksikan salah pada kelas non query dapat dilihat pada Tabel 6.15.

Berdasarkan pada Tabel 6.15, diketahui bahwa kesalahan dalam klasifikasi bukanlah kesalahan model dalam mengklasifikasikan namun, karena terdapat teks yang seharusnya termasuk ke dalam kelas non-query namun diberi label query, seperti pada no 468 dan 524. Untuk nomor 485, 486, dan 569 merupakan kelas non-query yang diprediksi kelas query. Untuk mengetahui kata yang menyebabkan model salah memprediksi maka pada Tabel 6.16 dilakukan perbandingan perubahan kata dari model lama dan model baru.

Tabel 6.14 Tabel perbandingan model lama dan baru

Teks		Label			Analisis
Sebelum dibakukan	Sesudah dibakukan	Pada Data yang seharusnya	Prediksi tanpa pembakuan	Prediksi dengan pembakuan	
Sorry sis, saya kurang suka. Kalo puncher yg kosong saya inden bisa?	maaf sis, saya kurang suka. kalau puncher yang kosong saya inden bisa?	Query	Query	Non-Query	Dengan dilakukan perubahan pada kata 'sorry' dirubah menjadi 'maaf' , kata 'yg' dirubah menjadi 'yang' membuat model salah memprediksi kelas menjadi kelas non query
Brp hari itu lesnya mba	berapa hari itu lesnya mbak	Query	Query	Non-Query	Dengan dilakukan perubahan pada kata 'brp' dirubah menjadi 'berapa' membuat model salah memprediksi kelas menjadi kelas non query

Cetakan norinya bkn 30rb mbk:D	Cetakan norinya bkn 30rb mbak:D	Query	Query	Non-Query	Dengan dilakukan perubahan pada kata 'mbk' dirubah menjadi 'mbak' membuat model salah memprediksi kelas menjadi kelas non query
Eh di pp saya ricemold kepala bunny ma bear	Eh di pp saya ricemold kepala bunny sama bear	Query	Query	Non-Query	Dengan dilakukan perubahan pada Kata 'ma' dirubah menjadi 'sama' membuat model salah memprediksi kelas menjadi kelas non query

Tabel 6.15 Teks yang diprediksi salah pada kelas non-query

No	Teks	Kelas pada model	Kelas pada data asli
468	Sama yang flower	Query	Non-query
485	<p>Ayo ketemu komunitas AYAHBUNDA bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREAST'</p> <p>Minggu - 8 Juli 2012</p> <p>Pk. 09.30 - 13.00</p> <p>WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya</p> <p>Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila</p> <p>Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda)</p> <p>Dress code : touch of gold</p> <p>50 peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan</p>	Query	Non-query

No	Teks	Kelas pada model	Kelas pada data asli
	bingkisan menarik. Ticket Box : ANGELINA - 0856 312 7797 Pin BB : 20EB14C0 Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -		
486	Ayo ketemu komunitas AYAHBUNDA bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREASI' Minggu - 8 Juli 2012 Pk. 09.30 - 13.00 WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda) Dress code : touch of gold 50	Query	Non-query

No	Teks	Kelas pada model	Kelas pada data asli
	peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan bingkisan menarik. Ticket Box : ANGELINA - 0856 312 7797 Pin BB : 20EB14C0 Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -		
524	Mbak puncher face yang murmer <80><c9><ec><d4><82> _<81><e4><ae><8d><ae><8d><e5><bb><80><fc>? :D	Query	Non-query
569	Kirimnya berapa paket?	Query	Non-query

Tabel 6.16 Tabel perbandingan model lama dan baru

No	Model lama	Model baru	Perubahan
468	Sama yg flower	Sama yang flower	Kata 'yg' dirubah menjadi 'yang'
485	Ayo ketemu komunitas AYAHBUND A bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREASI' Minggu - 8 Juli 2012 Pk. 09.30 - 13.00 WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo	Ayo ketemu komunitas AYAHBUNDA bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREASI' Minggu - 8 Juli 2012 Pk. 09.30 - 13.00 WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda)	Tidak dilakukan perubahan

No	Model lama	Model baru	Perubahan
	<p>masak - Chef Haryo Pramoe Moderator - Sari Nila</p> <p>Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda)</p> <p>Dress code : touch of gold 50 peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan bingkisan menarik.</p> <p>Ticket Box : ANGELINA - 0856 312 7797 Pin BB : 20EB14C0</p> <p>Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -</p>	<p>Dress code : touch of gold 50 peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan bingkisan menarik.</p> <p>Ticket Box : ANGELINA - 0856 312 7797 Pin BB : 20EB14C0</p> <p>Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -</p>	
486	Ayo ketemu	Ayo ketemu	Tidak

No	Model lama	Model baru	Perubahan
	<p>komunitas AYAHBUND A bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREASI' Minggu - 8 Juli 2012 Pk. 09.30 - 13.00 WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila Harga tiket : 75000 (dapat lunch dan</p>	<p>komunitas AYAHBUNDA bersama NUTRILON SOYA untuk anak anak Di seminar 'ATASI ALERGI , BEBAS BERKREASI' Minggu - 8 Juli 2012 Pk. 09.30 - 13.00 WIB Di Ballroom 1 dan 2 Hotel Sheraton Surabaya Jl. Embong Malang Surabaya Bersama : Dr. Anang Hendaryanto SpA (K) Dokter spesialis anak Disertai demo masak - Chef Haryo Pramoe Moderator - Sari Nila Harga tiket : 75000 (dapat lunch dan paket cantik dari Ayahbunda) Dress code : touch of gold 50 peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan bingkisan menarik.</p>	<p>dilakukan perubahan</p>

No	Model lama	Model baru	Perubahan
	<p>paket cantik dari Ayahbunda)</p> <p>Dress code : touch of gold</p> <p>50 peserta pertama yang datang sebelum pk. 09.30 akan mendapatkan bingkisan menarik.</p> <p>Ticket Box : ANGELINA - 0856 312 7797</p> <p>Pin BB : 20EB14C0</p> <p>Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -</p>	<p>Ticket Box : ANGELINA - 0856 312 7797</p> <p>Pin BB : 20EB14C0</p> <p>Pesan tiket sekarang , sebelum kehabisan :) - bantuin forward ya ke teman teman , sapa tau ada yang berminat datang -</p>	
524	<p>Mba puncher face yg murmer</p> <p><80><c9><ec><d4><82><d4><82><81><e4><ae><8d><ae><8d><e5><80><fc>? :D</p>	<p>Mbak puncher face yg murmer</p> <p><80><c9><ec><d4><82><d4><82><81><e4><ae><8d><ae><8d><e5><80><fc>? :D</p>	<p>Kata 'mba' dirubah menjadi 'mbak' dan kata 'yg' dirubah menjadi 'yang'</p>

No	Model lama	Model baru	Perubahan
	80><fc>? :D		
569	Kirimnya brp paket?	Kirimnya berapa paket?	Kata 'brp' dirubah menjadi 'berapa'

Berdasarkan Tabel 6.16 diketahui bahwa dengan dilakukan pembakuan kata, tidak mempengaruhi model dalam memprediksi kelas. Hal ini terlihat dari jumlah teks yang diprediksi salah pada model lama dan baru adalah sama. Dari hasil perbandingan teks yang termasuk ke dalam kelas query dan non-query maka dapat disimpulkan bahwa hasil pembakuan yang dilakukan tidak mempengaruhi nilai akurasi secara signifikan dan cenderung menurunkan nilai akurasi. Hal ini bisa terjadi karena pada svm, semakin banyak fitur akan semakin baik, sedangkan dengan dilakukannya pembakuan kata jumlah fitur semakin sedikit sehingga memungkinkan terjadinya penurunan nilai akurasi.

BAB VII

KESIMPULAN DAN SARAN

Pada bab ini dibahas mengenai kesimpulan dari semua proses yang telah dilakukan dan saran yang dapat diberikan untuk pengembangan yang lebih baik.

7.1 Kesimpulan

Berdasarkan hasil penelitian pada tugas akhir ini, maka didapatkan kesimpulan sebagai berikut :

1. Klasifikasi teks permintaan informasi produk menggunakan *Support Vector Machine* menghasilkan akurasi klasifikasi terbaik dengan menggunakan model skenario Uji Coba II yang menggunakan kernel linear dengan parameter *cost* sebesar $2^{-12.2}$.
2. Dari hasil klasifikasi yang dilakukan diketahui bahwa penggunaan *kernel linear* lebih bagus dalam mengklasifikasikan teks dibandingkan dengan penggunaan *kernel radial*.
3. Nilai Akurasi, presisi, *recall*, dan *F-Measure* dari hasil klasifikasi yang digunakan adalah sebesar 94.74%, 93.18%, 96.09%, dan 96.18%. Berdasarkan nilai tersebut dapat dikatakan bahwa model yang dibuat baik dalam mengklasifikasikan teks permintaan informasi produk.
4. Pembakuan kata pada *corpus* tidak memberikan dampak yang signifikan terhadap akurasi dan cenderung menurunkan nilai akurasi.
5. Dengan dilakukan penghapusan menggunakan *stopwords* tidak memberikan dampak yang signifikan terhadap akurasi dan cenderung menurunkan nilai akurasi.

7.2 Saran

Berdasarkan hasil penelitian pada tugas akhir ini, maka saran untuk penelitian selanjutnya adalah sebagai berikut :

1. Data yang digunakan bisa ditambahkan jumlahnya, agar data yang diproses menjadi lebih banyak sehingga bisa membuat model yang hasilnya lebih akurat.
2. Dalam pengaplikasian metode *grid search* untuk mencari akurasi model masih dilakukan secara manual, kedepannya dapat menggunakan *tools* lain yang bisa membuat proses pencarian menjadi lebih cepat dan optimal.
3. Dilakukan penambahan jenis *fitur* seperti *stemming* untuk meningkatkan akurasi.

DAFTAR PUSTAKA

- [1] Wyndo Mitra. (2016) StartupBisnis. [Online].
<http://startupbisnis.com/data-statistik-mengenai-pertumbuhan-pangsa-pasar-e-commerce-di-indonesia-saat-ini/>
- [2] Budi Santoso, Fajar Kurnia Dessyanto Boedi P, "Aplikasi Mobile Commerce Penjualan Buku (Studi Kasus Pada Penerbit PRO-U Media Yogyakarta," UPN "Veteran" Yogyakarta, Yogyakarta, 2010.
- [3] J.Simarmata, "Aplikasi Mobile Commerce Menggunakan PHP dan MySQL," Yogyakarta, 2006.
- [4] Romi Satria Wahono dan Abdul Syukur Abdul Razak Naufal, "Penerapan Bootstrapping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan ," Journal of Intelligent Systems , vol. 1 , December 2015.
- [5] Rizki Muliono, "Sidimpuan), Perancangan Web E-Commerce Jual Beli Batu Cincin Dengan Allgoritma Apriori (Studi Kasus Toko Batu Akik Murah Padang," Pelita Informatika Budi Darma, vol. VII, no. 3, Agustus 2014.
- [6] Agustinus Bimo Gumelar Cahyo Darujati, "Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia ," Jurnal Link, vol. 16, no. 2, February 2012.
- [7] Indah Fitri Astuti, Awang Harsa Kridalaksana Agus Setiawan, "Klasifikasi Dan Pencarian Buku

Referensi Akademik Menggunakan Metode Naive Bayes Classifier Studi Kasus : Perpustakaan Daerah Provinsi Kalimantan Timur," Jurnal Informatika Mulawarman , vol. 10, Februari 2015.

- [8] Acmad Nrhadi, "Klasifikasi Konten Berita Digital Bahasa Indonesia Menggunakan Support Vector Machines (SVM) Berbasis Particle Swarm Optimization (PSO)," Jurnal Bianglala Informatika , vol. 3, no. 2, September 2015.
- [9] Achmad Ridhok, Jendi Hardono Lailil Muflikha, "Klasifikasi Kondisi Penderita Penyakit Hepatitis Dengan Menggunakan Metode Support Vector Machine (SVM)," Universitas Brawijaya, 2013.
- [10] Hendri Murfi Djati Kerami, "Kajian Kemampuan Generalisasi Support Vector Machine Dalam Pengenalan Jenis Spice Sites Pada Barisan Dna ," Makara Sains, vol. 8, pp. 89-95, December 2004.
- [11] Agung Hardianto, M.Ridok Dewi Y.Liliana, "Indonesian News Classification using Support Vector Machine," World Academy of Science, Engineering and Technology, vol. 5, September 2011.
- [12] Mira Kania Sabariah, ST.,MT., Alfian Akbar Gozali, ST.,MT. Elly Susilowati, "Implementasi Metode Support Vector Machine Untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter," Teknik Informatika, Universitas Telkom, Bandung ,.
- [13] Chih-Chung Chang, and Chih-Jen Lin Chih-Wei Hsu, "A Practical Guide to Support Vector Classification," National Taiwan University, Taiwan, 2003.
- [14] Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard,

- Chih-Jen Lin Yin-Wen Chang, "Training and Testing Low-degree Polynomial Data Mappings via Linear SVM," *Journal of Machine Learning Research* 11, vol. 11, 2010.
- [15] Adyatma Bhaskara Hutomo, "Klasifikasi Dokumen Berita Menggunakan Metode Support Vector Machine Dengan Kernel Radial Baasis Function," Departemen Ilmu Komputer, Institut Pertanian Bogor, Bogor, 2014.
- [16] Hsuan-Tien Lin and Chih-Jen Lin, "A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods," Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003.
- [17] INNA SABILY KARIMA, "Optimasi Parameter Pada Support Vector Machine Untuk Klasifikasi Fragmen Metagenome Menggunakan Algoritme Genetika," Program Studi Ilmu Komputer, Institut Pertanian Bogor, Bogor, 2014.
- [18] Alan Prahutama, Tiani Wahyu Utami Hasbi Yasin, "Prediksi Harga Saham Menggunakan Support Vector Regression Dengan Algoritma Grid Search ," *Media Statistika* , vol. 7, pp. 29 - 35, June 2014.
- [19] Chin Wei Hsu, "A Practical Guide to Support Vector Classification," Computer Science, National Taiwan University , Taiwan, 2010.
- [20] C.Watters, M.Shepherd A.Basu, "Support Vector Machine for Text Categorization," in *Hawai International Conference on System Sciences*, Canada, 2002.

- [21] David Meyer. (2015, August) Package ‘e1071’. [Online]. <https://cran.r-project.org/web/packages/e1071/e1071.pdf>
- [22] Stephen Turner. (2015, February) Split a Data Frame into Testing and Training Sets in R. [Online]. <http://www.gettinggeneticsdone.com/2011/02/split-data-frame-into-testing-and.html>
- [23] Ingo Feinerer. (2015, July) Introduction to the tm Package Text Mining in R. [Online]. <https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf>
- [24] Matt Dowle Garrett Golemund. (2015, June) 15 Easy Solutions To Your Data Frame Problems In R. [Online]. <https://www.datacamp.com/community/tutorials/15-easy-solutions-data-frame-problems-r>
- [25] Garrett Golemund Matt Dowle. (2015, Maret) Machine Learning in R for beginners. [Online]. <https://www.datacamp.com/community/tutorials/machine-learning-in->
- [26] Ms. Snehlata S. Dongre, Dr. Latesh Malik Mr.Rushi Longadge, "Class Imbalance Problem in Data Mining: Review," *International Journal of Computer Science and Network*, vol. 2, no. 1, February 2013
- [27] Nurandi. (2013) github. [Online]. <https://raw.githubusercontent.com/nurandi/nurandi.net/master/data/stopwords-id.txt>

BIODATA PENULIS



Penulis lahir di Jakarta, 25 Juni 1994, dengan nama lengkap Dea Andia Rachmawati. Penulis merupakan anak kedua dari 2 bersaudara.

Riwayat pendidikan penulis yaitu TK Al-Azhar Jakapermai, SD Al-Azhar Jakapermai, SMP Labschool Jakarta, dan SMA Negeri 81 Jakarta, dan akhirnya penulis masuk menjadi mahasiswa Sistem Informasi angkatan 2012 melalui jalur SNMPTN dengan NRP

5212100177.

Selama kuliah penulis bergabung dalam organisasi kemahasiswaan, yaitu Himpunan Mahasiswa Sistem Informasi ITS dan Badan Eksekutif Mahasiswa ITS. Pada organisasi tersebut penulis mengikuti berbagai kegiatan dan menjadi Staff PSDM di BEM selama 1 tahun.

Di Jurusan Sistem Informasi penulis juga menjadi asisten dan mengambil bidang minat Akuisisi Data dan Diseminasi Informasi. Penulis dapat dihubungi melalui email deaandiaa49@gmail.com.